Microbial production and consumption of hydrocarbons in the global ocean

Connor R. Love[1*] (0000-0002-7801-3579), Eleanor C. Arrington[1*] (0000-0002-8078-396X), Kelsey M. Gosselin[1](0000-0001-5926-5415), Christopher M. Reddy[2](0000-0002-7814-2071), Benjamin A.S. Van Mooy[2] (0000-0002-2804-6508), Robert K. Nelson[2](0000-0003-0534-5801), David L. Valentine[3](0000-0001-5914-9107)

*These authors contributed equally.

[1]Interdepartmental Graduate Program for Marine Science, University of California – Santa Barbara, Lagoon Road, Santa Barbara, California 93106, United States
[2]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, 86 Water Street, Woods Hole, Massachusetts 02543, United States
[3]Department of Earth Science and Marine Science Institute, University of California – Santa Barbara, Lagoon Road, Santa Barbara, California 93106, United States

Corresponding author: David L. Valentine, Department of Earth Science, University of California, Webb Hall 2017, Santa Barbara, California 93106-9630, United States; +1-(805)-893-2973, valentine@ucsb.edu

ABSTRACT

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.

Introduction

Hydrocarbons are released into the ocean via natural oil seeps and industrial spills associated with extraction, transportation and consumption of oil, totaling ~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 (Fig. 1). In total, we quantified alkane concentration in 441 particulate samples (≥ 0.2 μm), mainly in triplicate (Methods, Supplementary Information Table 4, 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, Extended Data Fig. 4) and cultivation work (see 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 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 Fig. 2-3) and ^{13}C enrichment of pentadecane. Incubations were conducted shipboard at ambient temperature and light level (see Methods). In total, we quantified alkane production in 31 samples, from five of the seven stations, mainly in triplicate (see Supplementary Information Table 4). Pentadecane production varies between ~3-30 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 (see 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 penetrates (1% PAR) (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, 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 cell^{-1} d^{-1}) compared to 30% and 10% PAR (0.26 ± 0.05 and 0.13 ± 0.05 fg cell^{-1} d^{-1},

respectively) (Fig. 2b), indicating that cyanobacteria at or near the DCM produce more pentadecane per cell per unit 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 to 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 pentadecane 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 (Source Data Fig. 2) 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 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 is 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 (i.e., reservoir magnitude) and global production of pentadecane produced by *Prochlorococcus* and *Synechococcus* (i.e., turnover rate or input). Importantly, we assume consumption balances production (i.e., 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 (see 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 (Fig. 1 and Fig. 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 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 (see 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 pentadecane yr^{-1} by method 1 and

274-649 Tg pentadecane yr^{-1} by method 2 (Supplementary Information Table 1). By comparison, the total quantity of petroleum estimated to reach the ocean annually from all sources is 1.3 Tg[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].

In order 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 3 oligotrophic stations (see 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 days) – weighted slightly towards the slower environmental turnover of *Synechococcus* (1-6 days)[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, Extended Data Fig. 6). Blooms on pentadecane were dominated by *Alcanivorax* at station 6 and *Methylophaga* at station 3, whereas the addition of particles favored *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 to C_{11}) and long chain *n*-alkanes (C_{12}-C_{30}) – to be common among these taxa, with up to 10 copies (*almA* + *alkB*) per genome (Extended Data Fig. 9, Supplementary Information 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 subunit 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 (i.e., 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[35], we also analyzed 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 point to a need to analyze *rhdA* hits in close detail before assuming hydrocarbon oxidation functionality (see Supplementary Information 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 analyzed gene abundance from the Tara Oceans dataset. 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 Fig. 7-8). 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 and DCM stations (Extended Data Fig. 8, Supplementary Data 2). This highly successful group of surface-ocean dwelling archaea is known for a chemoorganoheterotrophic lifestyle targeting lipids, proteins, and amino acids[36,37] and can utilize photoheterotrophy, but a key role in biohydrocarbon cycling would be unexpected as MGII are not among the ~300 genera of bacteria and archaea[38,39] 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 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 to integrated production (1.76 mg nC$_{15}$ m^{-2} d^{-1}) from the same three stations, finding that only ~1×10^{-4} % of pentadecane production was exported below 150m. 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[40](Fig. 3a). Pentane was used for these experiments as it is a model non-biogenic compound that is unique to and abundant in petroleum, a structural analog to pentadecane (linear chain with an odd number of carbons), readily bioavailable based on its higher vapor pressure and aqueous solubility, relatively-low in toxicity, and is known to partition to the ocean's interior following release from the seafloor[42,43]. We find the microbial response to *n*-pentane to be structured by proximity to seepage, with ~9X 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 ~10X 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., (2015). We further demonstrate a decoupling between biological alkanes and petroleum-derived hydrocarbons that points to a complex interplay of chemical composition and biogeography that structure 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 harbor abundant Eukaryotic phytoplankton, many of which are capable of producing hydrocarbons[44–46]. Based on our qualitative observations from productive waters on the continental shelf of the Northwest Atlantic Ocean (see Supplementary Note), we expect such environments to harbor a dynamic and complex hydrocarbon cycle including biological alkanes and alkenes, structured in-part by proximity to continental sources and interaction with the sea floor. 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.

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

283 database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam31.0, Version 31.0) and the Ocean
284 Microbial Reference Gene Catalogue (http://ocean-microbiome.embl.de/companion.html).
285
286 **Code Availability Statement**
287 There is no custom code used.
288
289 **Main Text References**

290 1. National Research Council. *Oil in the Sea III*. (2003). doi:10.17226/10388
291 2. Han, J., McCarthy, E. D., Hoeven, W. V., Calvin, M. & Bradley, W. H. Organic
292 Geochemical Studies II. A Preliminary Report On The Distribution of Aliphatic
293 Hydrocarbons in Algae, in Bacteria, and in Recent Lake Sediment. *Proc. Natl. Acad. Sci.*
294 **59**, 29–33 (1968).
295 3. Li, X., del Cardayre, S. B., Popova, E., Schirmer, A. & Rude, M. A. Microbial Biosynthesis
296 of Alkanes. *Science (80-.).* **329**, 559–562 (2010).
297 4. Coates, R. C., Podell, S., Korobeynikov, A., Lapidus, A. & Pevzner, P. Characterization of
298 Cyanobacterial Hydrocarbon Composition and Distribution of Biosynthetic Pathways.
299 *PLoS One* **9**, 85140 (2014).
300 5. White, H. K. *et al.* Examining inputs of biogenic and oil-derived hydrocarbons in surface
301 waters following the deepwater horizon oil spill. *ACS Earth Sp. Chem.* **3**, 1329–1337
302 (2019).
303 6. Lea-Smith, D. J. *et al.* Contribution of cyanobacterial alkane production to the ocean
304 hydrocarbon cycle. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13591–13596 (2015).
305 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 (80-.).* **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. U. S. A.* **110**, 9824–9829 (2013).

10. Karl, D. M. & Church, M. J. Microbial oceanography and the Hawaii Ocean Time-series programme. *Nature Reviews Microbiology* **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. Res. Part I-Oceanographic Res. Pap.* **48**, 2373–2395 (2001).

14. Johnson, Z. I. *et al.* Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science (80-.).* **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. Part A, Oceanogr. Res. Pap.* **36**, 1621–1634 (1989).

16. Karl, D. M. & Church, M. J. Ecosystem Structure and Dynamics in the North Pacific

329 Subtropical Gyre: New Views of an Old Ocean. *Ecosystems* **20**, 433–457 (2017).
330 17. Knoot, C. J. & Pakrasi, H. B. Diverse hydrocarbon biosynthetic enzymes can substitute for
331 olefin synthase in the cyanobacterium Synechococcus sp. PCC 7002. *Sci. Rep.* **9**, (2019).
332 18. Martiny, A. C., Kathuria, S. & Berube, P. M. Widespread metabolic potential for nitrite and
333 nitrate assimilation among Prochlorococcus ecotypes. *Proc. Natl. Acad. Sci. U. S. A.* **106**,
334 10787–92 (2009).
335 19. Saunois, M. *et al.* The global methane budget 2000-2017. *Earth Syst. Sci. Data* **12**, 1561–
336 1623 (2020).
337 20. Guenther, A. B. *et al.* The model of emissions of gases and aerosols from nature version 2.1
338 (MEGAN2.1): An extended and updated framework for modeling biogenic emissions.
339 *Geosci. Model Dev.* **5**, 1471–1492 (2012).
340 21. McGenity, T. J., Crombie, A. T. & Murrell, J. C. Microbial cycling of isoprene, the most
341 abundantly produced biological volatile organic compound on Earth. *ISME J.* **12**, 931–941
342 (2018).
343 22. Vaulot, D., Marie, D., Olson, R. J. & Chisholm, S. W. Growth of Prochlorococcus, a
344 photosynthetic prokaryote, in the equatorial Pacific Ocean. *Science (80-.).* **268**, 1480–1482
345 (1995).
346 23. Mann, E. L. & Chisholm, S. W. Iron limits the cell division rate of Prochlorococcus in the
347 eastern equatorial Pacific. *Limnol. Oceanogr.* **45**, 1067–1076 (2000).
348 24. Zubkov, M. V. Faster growth of the major prokaryotic versus eukaryotic CO2 fixers in the
349 oligotrophic ocean. *Nat. Commun.* **5**, (2014).
350 25. Liu, H. B., Campbell, L. & Landry, M. R. Growth and mortality rates of Prochlorococcus
351 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. U. S. A.* **109**, 20229–20234 (2012).

28. Head, I. M., Jones, D. M. & Röling, W. F. M. Marine microorganisms make a meal of oil. *Nature reviews. Microbiology* **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 and 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. 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).

37. Tully, B. J. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat. Commun.* **10**, (2019).

38. Hazen, T. C., Prince, R. C. & Mahmoudi, N. Marine Oil Biodegradation. *Environ. Sci. Technol.* **50**, 2121–2129 (2016).

39. Prince, R. C., Amande, T. J. & McGenity, T. J. Prokaryotic Hydrocarbon Degraders. in *Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes* (ed. McGenity, T. J.) 1–39 (Springer International Publishing, 2019). doi:10.1007/978-3-030-14796-9_15

40. MacDonald, I. R. *et al.* Natural and unnatural oil slicks in the Gulf of Mexico. *J. Geophys. Res. Ocean.* **120**, 8364–8380 (2015).

41. Valentine, D. L. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol.* **5**, 316–323 (2007).

42. 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**, (2011).

43. Ryerson, T. B. *et al.* Chemical data quantify Deepwater Horizon hydrocarbon flow rate and environmental distribution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 20246–20253 (2012).

44. Sorigué, D. *et al.* Microalgae synthesize hydrocarbons from long-chain fatty acids via a light-dependent pathway. *Plant Physiol.* **171**, 2393–2405 (2016).

45. Sorigué, D. *et al.* An algal photoenzyme converts fatty acids to hydrocarbons. *Science (80-.).* **357**, 903–907 (2017).

46. Aleksenko, V. A. *et al.* Phylogeny and structure of fatty acid photodecarboxylases and glucose-methanol-choline oxidoreductases. *Catalysts* **10**, 1–21 (2020).

Corresponding author: David L. Valentine, Department of Earth Science, University of California, Webb Hall 2017, Santa Barbara, California 93106-9630, United States; +1-(805)-893-2973, valentine@ucsb.edu

Acknowledgements

We thank Georges Paradis for analytical support; the *R/V Neil Armstrong* captain and crew for support at sea; the Hawai'ian Oceanographic Time Series program and the crew of the R/V *Kilo Moana* for enabling a preliminary study; John Hayes and Penny Chisholm for their advice on the study, Justin Ossolinski, Helen Fredricks, Bethany Jenkins and Robert Swarthout for assistance on the *R/V Armstrong* cruise; Tara McKinnon for assistance in dissolved inorganic carbon isotope measurements; Alina Ebling for total particulate phosphate measurements; Nicole Poulton for flow cytometry analysis; and Katherine 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 number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This project was supported by NSF grants OCE-1635562, OCE-1536346, OCE-1756254, and OCE-1634478.

Author Contributions

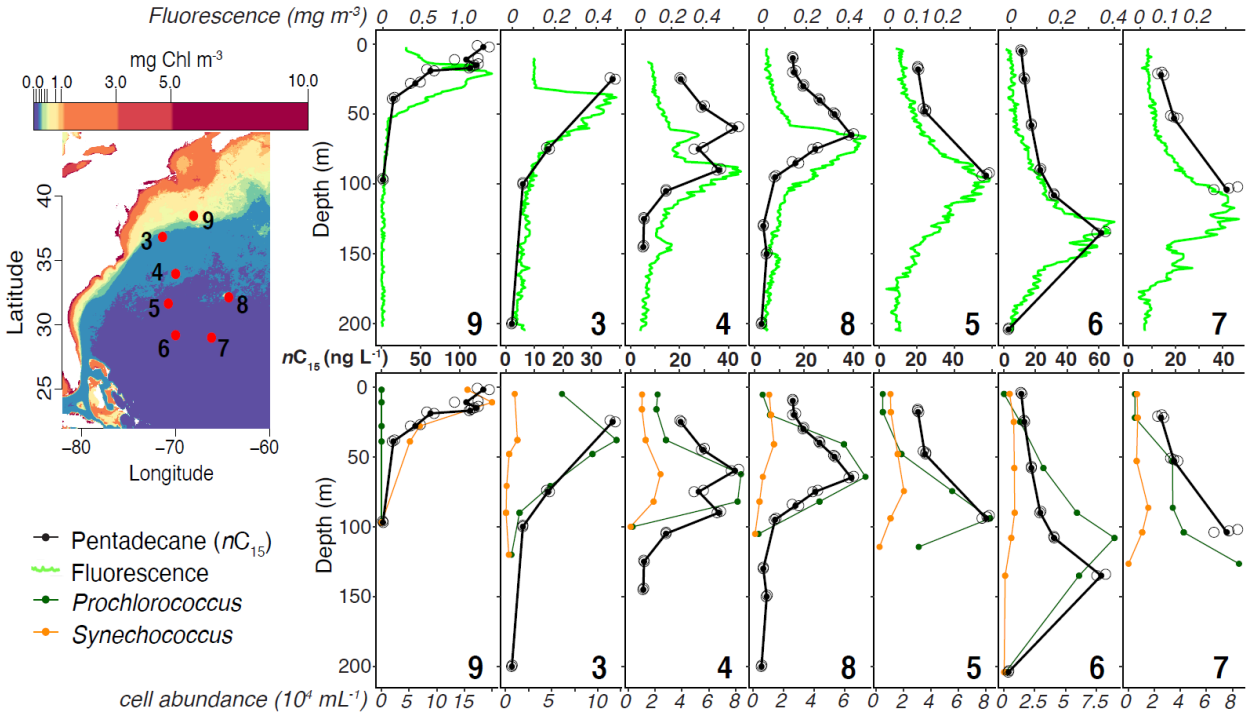
C.L. and E.A. contributed equally, with C.L. contributing the organization of field experiments and measurements of pentadecane production and concentration and corresponding data analysis, E.A. contributed biodegradation experiments and bioinformatic work. B.V.M contributed nutrient data, cell count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

Competing Interest Declaration

There are no competing interests.

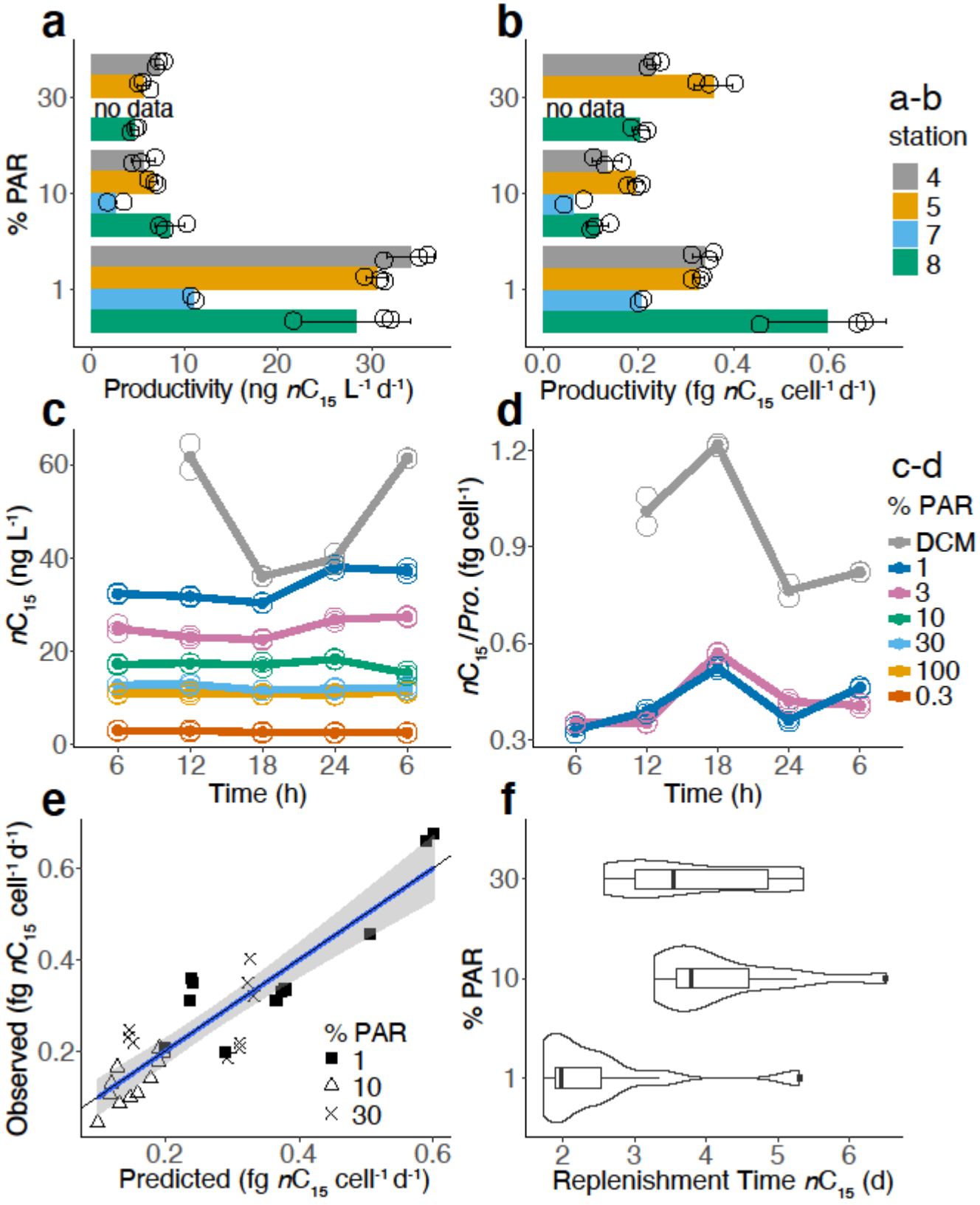
Additional Information

Supplementary Information is available for this paper. Correspondence and request for materials should be addressed to David Valentine.



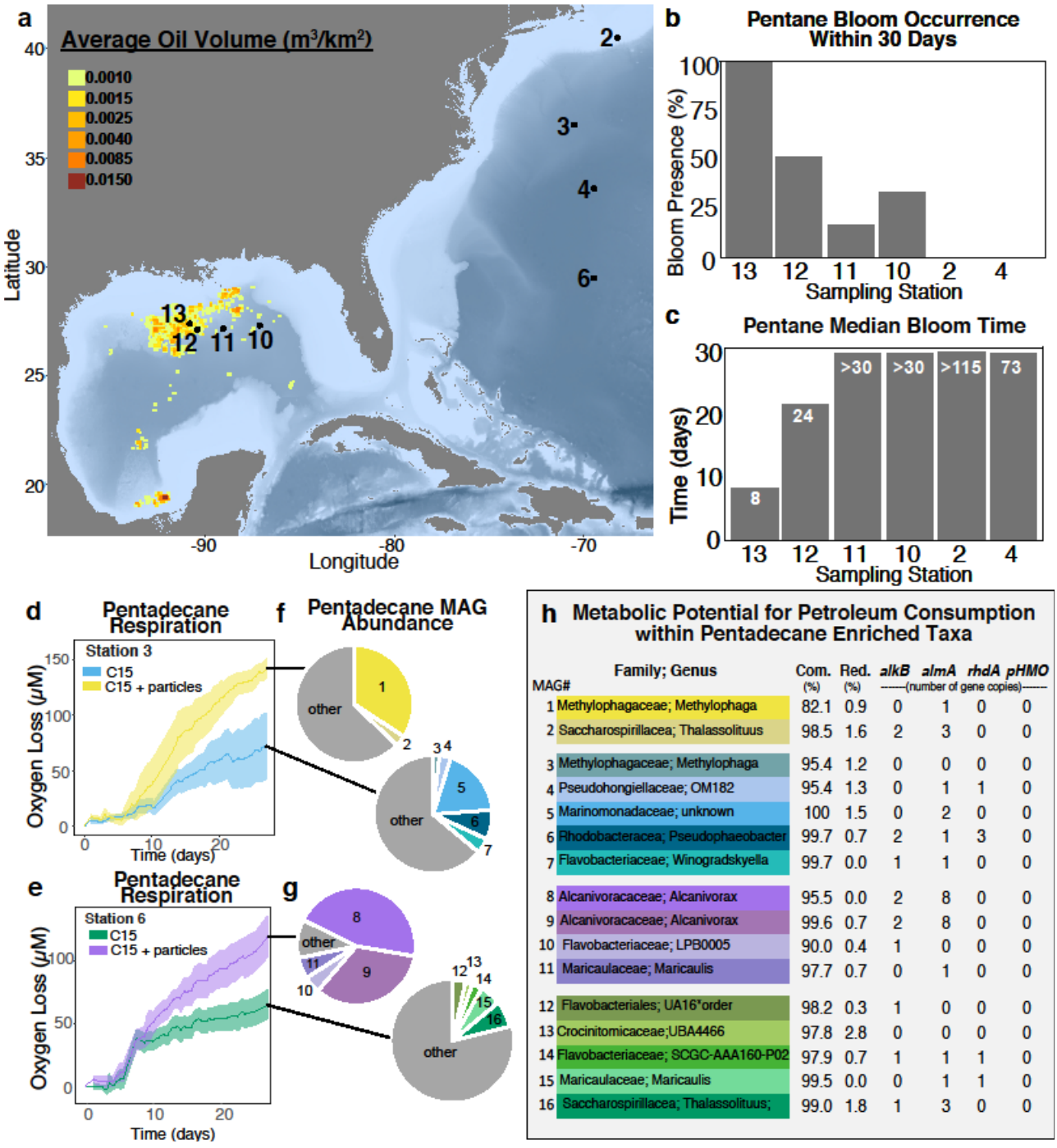
433

434 **Figure 1.**



435

436 **Figure 2.**



MAG#	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

Figure 3.

Figure Legends

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 chlorophyll 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 represents the contents of one distinct sample bottle (see Methods). Replicates are sequentially moved 1-meter 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 (see Supplementary Information Table 4).

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**-**b** Volumetric and cellular (cell = *Pro.* + *Syn.*) pentadecane production were calculated using pentadecane concentration and ^{13}C enrichment from incubation experiments (see Methods). Data displayed as open black circles with bar representing mean production rate, error bars show standard deviation for n = 3. **c**-**d** Diel change in pentadecane concentration and pentadecane per *Prochlorococcus* cell show the lower euphotic zone and particularly the Deep Chlorophyll Maximum (DCM) is most dynamic (see Extended Data Fig. 4-5), data are plotted as open circles, mean 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), gray 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 (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x 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.

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[40] with increasing distance from intense seepage as follows 13, 12, 11, 10, 6, 4, 3, and 2. **b-c** Occurrence and timing of respiratory blooms on pentane (petroleum proxy compound) at 1000 m with increasing distance from seepage (n = 6 at station 13-10, n = 8 at station 2 and 4). Experiment duration 30 days at stations 10-13 and 115 days for stations 2 and 4. Numbers listed in white are median bloom times over the entire duration of the experiment which varied from 30 days to 115 days, assuming all bottles bloom given enough time. **d-e** Pentadecane respiration at 500 m +/- particles (solid line indicates mean, shading indicates ± SD, n = 6 at station 3 and n=4 at station 6). **f-g** Relative abundance of metagenomes (MAGs) at final time point of pentadecane incubations (~28 days). **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 subunit A, B, and C). *MAG 12 is unclassified at the family and genus level therefore we have listed the class and order. For panels **b-e**, "n" describes the number of biologically independent incubations.

Online 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 AM) for all sampling except for the diel experiment. Salinity, density, temperature, fluorescence and percent photosynthetically active radiation (% 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 six-hour intervals through a full 24-hour 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, the entire CTD rosette was 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 (i.e. smoking and painting). Control samples were collected by pouring clean MilliQ water into the Niskin bottles and letting it sit for 30 minutes and then filtering the water using the same procedure for all samples. No pentadecane of considerable quantity (> 2 ng L^{-1}) was found in control samples and thus validated 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 hr 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 Inc., ^{13}C 99%) was added to the 2L polycarbonate bottles to achieve a 480 ‰ enrichment in seawater DIC. Dark control bottles were covered completely beforehand with aluminum foil before tracer addition and kill control bottles were treated with Zinc Chloride to 2% $ZnCl_2$ (m/v) before tracer addition. 2 L bottles were then 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[47] and photosynthetically active radiation attenuation by the bags was quantified using a spherical quantum sensor (Licor). Bottles were harvested at 0-hour (initial), 5, 10, 20 and 30-hour (final) time points for the 30% PAR light bags and at t = 0 hour and t = 30 hour final for the 10% and 1% light levels, care was taken to reduce light exposure in the ship-board laboratory when preparing for incubation by placing bottles into covered tubs. A 2 mL aliquot was taken for ^{13}C-DIC prior to filtration. Filters were placed into pre-combusted aluminum foil packets and immediately frozen at -20° C for later analysis.

A preliminary culture experiment was conducted to assess the percent of all cyanobacterial hydrocarbons within membranes, i.e. what percent 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., 2016 and the low solubility of straight chain hydrocarbons 15-17 carbons in length.

Hydrocarbon Extraction and Analysis

A modified Bligh-Dyer[48] 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 amounts of each solvent was used according to Van Mooy et al. 2008 and a 10-minute 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 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 pre-combusted 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 x 0.25 mm 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 at 70 °C held for 2 minutes, 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 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, GCxGC-FID and GCxGC-TOF (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, Saint Joseph, Michigan). 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 I.D., 0.25 μm df) and second-dimension separations were performed on a 50% phenyl polysilphenylene-siloxane column (SGE BPX50, 1.2-m length, 0.10 mm I.D., 0.1 μm df). The temperature program of the main oven was held isothermal at 50 °C (15 min) and was then ramped from 50 to 335 °C at a rate of 1.5 °C min^{-1}. The second-dimension oven was isothermal at 60 °C (15 min) and then ramped from 60 to 345 °C at a rate of 1.5 °C min^{-1}. The hot jet pulse width was 0.75 seconds, while the modulation period

between stages was 7.50 seconds and a 3.00 seconds cooling period, for the FID method, and 10.00 seconds and a 4.25 second cooling period for the TOF method. FID data was sampled at an acquisition rate of 100 data points per second, while the TOF data was sampled at an acquisition rate of 50 spectra per second in the mass range of 40 to 500 atomic mass units (amu). 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 Dissolved Inorganic Carbon Isotope Measurements

Compound-specific isotope analysis was performed after concentration analysis on a gas chromatograph combustion isotope ratio mass spectrometer (GC/C-IRMS) with a Trace GC (Thermo Finnigan) set up to a GC-C/TC III (Finnigan^{TM}) 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. Temperature ramp was conducted starting at 70 °C and held for 2 minutes, then a 3 °C min^{-1} ramp to 120 °C, hold for 0 minutes, then a 6 °C min^{-1} ramp to 185 °C, hold for 0 minutes then a 120 °C min^{-1} ramp to 290 °C, hold for 3 minutes. Inlet temperature was 260 °C, flow rate was held at 2.2 mL He min^{-1} with a splitless injection held for 0.5 minutes 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 *n*C_{15}, *n*C_{16} and *n*C_{17} at ~1.2 ng μL^{-1} and was 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.

Dissolved inorganic carbon ^{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 were followed closely to 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 1000 m. Water from the CTD Niskin bottles were transferred to 250 mL glass serum vials using a small length of Tygon tubing. Vials were filled for 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 prior to 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 station 3 and 6 were deployed for 24 hours at 150 m. For each particle addition, 10 mL of particle laden seawater was vortexed lightly for 1 minute, 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 hours with a fiber 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 loss 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 (see Methods). For information on 16S rRNA amplification, metagenomic reconstruction, and bioinformatic analysis of Tara Oceans dataset see Methods.

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-hour difference), these casts were all at ~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., 2010[49]. Samples were fixed with paraformaldehyde (0.5% final concentration) and stored at ~4 °C for 1-2 hours before long term storage in liquid nitrogen. An Influx cytometer was used with a 488 nm blue excitation laser, appropriate Chl-a (692 ± 20 nm) and phycoerythrin (585 ± 15 nm) bandpass filters, and was calibrated daily with 3.46 μm Rainbow Beads (Spherotech Inc. Lake Forest, Illinois, USA). Each sample was run for 4–6 min (~0.2–0.3 ml total volume analyzed), with log-amplified Chl-a and phycoerythrin fluorescence, and forward and right-angle scatter signals recorded. Data files were analyzed from two-dimensional scatter plots based on red or orange fluorescence and characteristic light scattering properties [50] using FlowJo 9.8 Software

650 (Becton Dickinson, San Jose, CA). Pico-autotrophs were identified as either *Synechococcus* or
651 *Prochlorococcus*, pico-eukaryotes or nano-eukaryotes based upon cell size and the presence or
652 absence of phycoerythrin. Nutrients were analyzed by the University of Washington Marine
653 Chemistry Laboratory.

654 **Calculations and Analyses, Statistics and Reproducibility**

655 All statistics and points within figures were conducted with distinct samples (not replicated
656 measurements of the same sample). Pentadecane production from compound-specific isotope
657 enrichment measurements were calculated using a published equation[51]. The time duration used in
658 the equation was from complete set up of the incubation to completion of filtering the water
659 through the filter. The value used for ^{13}C-DIC was the average of the whole dataset ($\delta^{13}C = 480‰$)
660 and the value used for unlabeled pentadecane was from a non-enriched sample ($\delta^{13}C = -20‰$)
661 because of variations in the time zero values from a slight but inevitable enrichment when bottles
662 were filtered in the laboratory (roughly one hour to filter the whole bottle in a well-illuminated
663 laboratory space).

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

669

670 **Quantification of global stock and production for cyanobacterial alkanes: Method 1**

671 Method 1 draws from direct observations of water column pentadecane stock and
672 production rates encountered in the North Atlantic subtropical gyre. We integrated the depth

673 profiles of pentadecane concentration for stations 4, 6 and 8 to calculate a mean water column
674 integrated stock of pentadecane with standard deviation and further integrated primary production
675 rates of pentadecane for stations 4, 5 and 8 from our isotope enrichment incubation experiments,
676 to obtain a mean water column production rate with standard deviation. Calculation of pentadecane
677 stock results in an average water column integrated stock of pentadecane of 3.42 ± 0.83 mg m^{-2},
678 and when scaled by the mean areal extent of the oligotrophic ocean (estimated at 204 × 10^6 km^2)
679 results in a standing stock of 0.70 ± 0.17 Tg (Supplementary Information Table 1). Calculation of
680 pentadecane production rate results in 1.76 ± 0.17 mg pentadecane m^{-2} d^{-1}, which multiplied by the
681 areal extent of the oligotrophic ocean yields 131 ± 13 Tg pentadecane per year (Supplementary
682 Information Table 1).

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

692 To integrate pentadecane production rate throughout the water column we used "typical"
693 oligotrophic stations that had production measurements at 30%, 10% and 1% PAR (stations 4, 5
694 and 8). All three stations had a very similar trend in productivity (Fig. 2A). We integrated by taking
695 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 modeled global populations of *Prochlorococcus* ($2.9 \pm 0.1 \times 10^{27}$) and *Synechococcus* ($7.0 \pm 0.3 \times 10^{26}$)[9] to scale our estimates for a global stock and utilized known doubling rates (1-2 days for *Prochlorococcus*, 1-6 days 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 cell^{-1} for *Prochlorococcus* and 0.60 ± 0.35 fg cell^{-1} 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/cell) and are slightly higher than reported of four strains of *Synechococcus* (0.25 ± 0.04 fg/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 *Synechococcus* to be 0.42 ± 0.25 Tg, for a total of 1.78 ± 1.24 Tg. See Supplementary Information Table 1 for estimates and comparisons to Lea-Smith et al. 2015.

DNA extraction

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

723 **16S rRNA gene amplification and analysis**

724 We amplified and barcoded the V4 region of the 16S rRNA gene using the method
725 described previously[52] with small modifications to the 16Sf and 16Sr primers[53]. Amplicon PCR
726 reactions contained 1 μL of template DNA, 2 μL of forward primer, 2 μL of reverse primer, and
727 17 μL of AccuPrime Pfx SuperMix. Thermocycling conditions consisted of 95° 2 min, 30 cycles
728 of 95°C for 20 secs, 55°C for 15 secs, 72°C for 5 min, and a final elongation at 72°C for 10 min.
729 Sample DNA concentrations were normalized using the SequelPrep Normalization Kit, cleaned
730 using the DNA Clean and Concentrator kit, visualized on an Agilent Tapestation, and quantified
731 using a Qubit fluorometer. Samples were sequenced and demultiplexed at the UC Davis Genome
732 Center on the Illumina MiSeq platform with 250nt, paired end reads. A PCR-grade water sample
733 was included in extraction, amplification, and sequencing as negative control to assess for DNA
734 contamination.

735 Trimmed fastq files were quality filtered using the fastqPairedFilter command within the
736 dada2 R package, version 1.9.3[54] with following parameters: truncLen=c(190,190), maxN=0,
737 maxEE=c(2,2), truncQ=2, rm.phix=TRUE, compress=TRUE, multithread=TRUE. Quality
738 filtered reads were dereplicated using derepFastq command. Paired dereplicated fastq files were
739 joined using the mergePairs function with the default parameters. Single Nucleotide Variant (SNV)

table was constructed with the makeSequenceTable command and potential chimeras were removed denovo using removeBimeraDenovo. Taxonomic assignment of the sequences was done with the assignTaxonomy command using the Silva taxonomic training dataset formatted for DADA2 v132. If sequences were not assigned, they were left as NA.

Metagenome assembly, binning, and relative abundance calculation

Metagenomic library preparation and shotgun sequencing were conducted at the UC Davis DNA Technologies Core. DNA was sequenced on the Illumina HiSeq4000 platform, producing 150bp paired-end reads with a targeted insert size of 400bp. Quality control and adapter removal was performed with Trimmomatic[55] (Version 0.36; parameters: leading 10, trailing 10, sliding window of 4, quality score of 25, minimum length 151bp) and Sickle[56] (Version 1.33 with paired end and sanger parameters). Concatenation of high-quality reads for replicate samples (for coassembly) was conducted prior to assembly (see Supplementary Information Table 2 below for more details on coassembly). The trimmed high-quality reads were assembled using metaSPAdes[57] (Version 3.8.1; parameters k= 21,33, 55, 77, 88, 127). The quality of assemblies was determined using QUAST[58] (Version 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[59] (Version 2.3.4.1; default parameters) with Samtools[60] (Version 1.7). Contigs greater than 2,500 bp were manually binned using Anvi'o with Centrifuge[61,62](Version 1.0.1) based on coverage uniformity (Version 5). Quality metrics for metagenome assembled genomes (MAGs) were determined using CheckM[63] (Version 1.0.7; default parameters). The taxonomic identity of each MAG was determined using GTDB-Tk[64] (Version 1.0.2) against The Genome Taxonomy Database[65]

762 (https://data.ace.uq.edu.au/public/gtdb/data/releases/release89/89.0/, Version r89). The length
763 normalized relative abundance of MAGs were determined for each sample as in Tully et. al 2018[66].

764 **Metagenomic annotation of hydrocarbon degradation genes**

765 Open reading frames were predicted for MAGs using Prodigal[67] (Version 2.6.3; default
766 parameters). Functional annotation was determined using HMMER3[68] (Version 3.1b2) against the
767 Pfam database[69] (Version 31.0) with an e-value cutoff of $1x10^{-7}$ and KofamScan (version 1.1.0)[70]
768 against the HMM profiles for KEGG/KO with a score cutoff of $1x10^{-7}$. For Figure 3H, to find
769 number of hits for *almA* we used Pfam (PF00743), for *rhdA* we used Pfam (PF00848), for *pHMO*
770 we summed Pfam hits (subunit a: PF02461, subunit b: PF04744, subunit c: PF04896).

771 For alkane-1-monooxygenase (*alkB*) we used both HMMER3 against the Pfam database
772 (PF00487) and KofamScan against the KEGG/KO hmm profiles (K00496). Each hit was
773 manually curated using Geneious Prime 2019.2.3 (https://www.geneious.com) to search for the
774 eight-histidine residues considered catalytically essential for function[71]. The base seed alignments
775 for both PF00487 and K00496 include the ancestrally related protein, fatty acid desaturase,
776 therefore we found it necessary to phylogenetically analyze each hit to determine relation to *alkB*
777 or fatty acid desaturase. Through this method we learned that the HMMER3 method with PfamID
778 PF00487 identifies more hits within each MAG for *alkB* than KofamScan with K00496; however,
779 those additional hits were generally more closely related to fatty acid desaturases than *alkB*.
780 Furthermore, the phylogenies produced are necessary to determine similarity to the related xylene
781 monooxygenase protein which acts on the methyl groups of xylene. We excluded any hits that
782 formed a well-supported monophyletic clade with xylene monooxygenase from our final number
783 of copies of *alkB*. In total, we used KofamScan with K00496 to search for *alkB*, manually curated
784 the results to ensure presence of eight-histidine residues essential for function, and

phylogenetically analyzed each hit for relation to *alkB* compared to fatty acid desaturase and xylene monooxygenase.

Phylogenetic analyses

Each putative *alkB* hit was aligned using MUSCLE[72] (Version 3.8.425). For Extended Data Figure 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[73]. All columns with >95% gaps were removed using TrimAL[74] (Version 1.2). Phylogenetic analysis of concatenated *alkB* was inferred by RAxML[75] (Version 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[76] (Version 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[77] 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[77]. The abundance of select genes involved in hydrocarbon oxidation were analyzed

807 from the Tara Ocean dataset. The total abundance of OM-RGC sequences matching the reference
808 gene identifier was normalized to the total abundance of the single copy gene *recA* (KEGG
809 identifier: K03553), as performed in previous studies, in order to calculate abundance on a per-
810 genome level[78,79]. The resulting data is included in Supplementary Data 1.

811 For select Tara Oceans stations we conducted further analysis of alkane-1-monooxygenase
812 to assess the diversity and abundance of the gene for oceanographic settings underlying the North
813 Atlantic subtropical gyre (Extended Data Figure 7-8). First, we took the assembled Tara Oceans[77]
814 data (see http://ocean-microbiome.embl.de/companion.html) and used Prodigal[67] (Version 2.6.3;
815 default parameters) to identify open reading frames. The resulting protein sequences were scanned
816 for *alkB* using the above method (KofamScan for K00496, manual curation for eight-histidine
817 residues and phylogenetically analyzed). Each curated hit was assigned a taxonomic classification
818 through homology search using BLAST[80] (Version 2.7.1) against the nr (version 38 accessed
819 December 2019) (see resulting data in Supplementary Data 2). Read mapping of high quality reads
820 from each respective station using Bowtie2[59] (Version 2.3.4.1) was used to determine the
821 abundance of each unique *alkB*-like protein at each station.

822 **Methods references**

823 47. Reshkin, S. J. & Knauer, G. A. Light stimulation of phosphate uptake in natural assemblages
824 of phytoplankton. *Limnol. Oceanogr.* **24**, 1121–1124 (1979).
825 48. Van Mooy, B. A. S., Moutin, T., Duhamel, S., Rimmelin, P. & Van Wambeke, F.
826 Phospholipid synthesis rates in the eastern subtropical South Pacific Ocean. *Biogeosciences*
827 **5**, 133–139 (2008).
828 49. Lomas, M. W. *et al.* Increased ocean carbon export in the Sargasso Sea linked to climate

829 variability is countered by its enhanced mesopelagic attenuation. *Biogeosciences* **7**, 57–70
830 (2010).
831 50. Durand, M. D. & Olson, R. J. Contributions of phytoplankton light scattering and cell
832 concentration changes to diel variations in beam attenuation in the equatorial pacific from
833 flow cytometric measurements of pico-, ultra and nanoplankton. *Deep. Res. Part II Top.*
834 *Stud. Oceanogr.* **43**, 891–906 (1996).
835 51. López-Sandoval, D. C., Delgado-Huertas, A. & Agustí, S. The 13 C method as a robust
836 alternative to 14 C-based measurements of primary productivity in the Mediterranean Sea.
837 *J. Plankton Res.* **40**, 544–554 (2018).
838 52. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development
839 of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence
840 data on the miseq illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120
841 (2013).
842 53. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina
843 HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
844 54. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon
845 data. *Nat. Methods* **13**, 581–583 (2016).
846 55. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina
847 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
848 56. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for
849 FastQ files (Version 1.33) [Software]. *Available at https://github.com/najoshi/sickle.* 2011
850 (2011).
851 57. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile

metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

58. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).

59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

60. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).

62. Eren, A. M. *et al.* Anvi'o: An advanced analysis and visualization platformfor 'omics data. *PeerJ* **2015**, e1319 (2015).

63. 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).

64. 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).

65. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 1–8 (2020). doi:10.1038/s41587-020-0501-8

66. 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).

67. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site

identification. *BMC Bioinformatics* **11**, (2010).

68. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, 1002195 (2011).

69. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

70. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

71. 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).

72. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

73. 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).

74. 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).

75. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

76. Rambaut, A. FigTree v1. 4. (2012).

77. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* **348**, (2015).

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

898 pathway for phosphonate degradation. *Environ. Microbiol.* **21**, 2402–2414 (2019).
899 79. Martinez, A., Tyson, G. W. & Delong, E. F. Widespread known and novel phosphonate
900 utilization pathways in marine bacteria revealed by functional screening and metagenomic
901 analyses. *Environ. Microbiol.* **12**, 222–238 (2010).
902 80. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**,
903 (2009).

Supplementary Information for

Microbial production and consumption of hydrocarbons in the global ocean

Connor R. Love[1*] (0000-0002-7801-3579), Eleanor C. Arrington[1*] (0000-0002-8078-396X), Kelsey M. Gosselin[1](0000-0001-5926-5415), Christopher M. Reddy[2](0000-0002-7814-2071), Benjamin A.S. Van Mooy[2] (0000-0002-2804-6508), Robert K. Nelson[2](0000-0003-0534-5801), David L. Valentine[3](0000-0001-5914-9107)

*These authors contributed equally.

[1]Interdepartmental Graduate Program for Marine Science, University of California – Santa Barbara, Lagoon Road, Santa Barbara, California 93106, United States
[2]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, 86 Water Street, Woods Hole, Massachusetts 02543, United States
[3]Department of Earth Science and Marine Science Institute, University of California – Santa Barbara, Lagoon Road, Santa Barbara, California 93106, United States

Corresponding author: David L. Valentine, Department of Earth Science, University of California, Webb Hall 2017, Santa Barbara, California 93106-9630, United States; +1-(805)-893-2973, valentine@ucsb.edu

Supplementary Tables

	Method 1	Method 2	Lea-smith et al., 2015
Stock (Tg *n*C15)	0.70 ± 0.17	1.78 ± 1.24	1.59
Production (Tg *n*C15 yr^{-1})	131 ± 13	274-649	270-583
***n*C_{15} Consuming Cells (30% PAR)**	20-200 cells mL^{-1}		
***n*C_{15} Consuming Cells (1% PAR)**	100-1000 cells mL^{-1}		

Supplementary Information Table 1: Global reservoirs and fluxes of pentadecane and estimates of a supported bacterial or archaeal community relying solely on pentadecane. We assume a conversion efficiency range (pentadecane to biomass) of 5-50%, a dry carbon mass of hydrocarbon degrading cells of 120 fg C/cell[1], and a cellular turnover rate of 0.1 d^{-1} (see Supplementary Note for additional details).

MAG name	Assembly Method (sample ID)	Comp. (%)	Red. (%)	# Contigs	N50 (bp)	GC Content	Length (bp)	Taxonomy
1_P54_P55_methylophagaceae	Coassembly (54 + 55)	82.1	0.9	211	25666	45.2	2556684	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Nitrosococcales;f Methylophagaceae;g Methylophaga;s Methylophaga sp002696735
2_P54_P55_saccharospirillaceae	Coassembly (54 + 55)	98.5	1.6	45	126959	46.7	3730513	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Saccharospirillaceae;g Thalassolituus;s Thalassolituus oleivorans
3_P61_P62_methylophagaceae	Coassembly (61 + 62)	95.4	1.2	94	44824	45.8	2556822	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Nitrosococcales;f Methylophagaceae;g Methylophaga;s Methylophaga thiooxydans
4_P61_P62_pseudohongiellaceae	Coassembly (61 + 62)	95.4	1.3	74	79032	53.3	2910293	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Pseudohongiellaceae;g OM182;s OM182 sp001438145
5_P61_P62_marinomonadaceae	Coassembly (61 + 62)	100	1.5	17	479598	42.5	4846406	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Marinomonadaceae;g ;s unknown
6_P61_P62_rhodobacteraceae	Coassembly (61 + 62)	99.7	0.7	40	243772	59.1	4474503	d Bacteria;p Proteobacteria;c Alphaproteobacteria;o Rhodobacterales;f Rhodobacteraceae;g Pseudophaeobacter;s unknown
7_P61_P62_flavobacteriaceae	Coassembly (61 + 62)	99.7	0.0	23	267233	33.0	2816890	d Bacteria;p Bacteroidota;c Bacteroidia;o Flavobacteriales;f Flavobacteriaceae;g Winogradskyella;s Winogradskyella sp002163855
8_P98_alcanivoracaceae	Single assembly (98)	95.5	0.0	25	422209	58.3	3726153	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Alcanivoracaceae;g Alcanivorax;s Alcanivorax sp000155615
9_P100_alcanivoracaceae	Single assembly (100)	99.6	0.7	27	1001868	58.0	4606115	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Alcanivoracaceae;g Alcanivorax;s Alcanivorax sp000155615
10_P100_P98_flavobacteriaceae	Coassembly (98 + 100)	90.0	0.4	64	56874	40.5	2497634	d Bacteria;p Bacteroidota;c Bacteroidia;o Flavobacteriales;f Flavobacteriaceae;g LPB0005;s unknown
11_P100_P98_maricaulaceae	Coassembly (98 + 100)	97.7	0.7	44	108979	64.1	3165928	d Bacteria;p Proteobacteria;c Alphaproteobacteria;o Caulobacterales;f Maricaulaceae;g Maricaulis;s unknown
12_P102_P105_flavobacteriales	Coassembly (102 + 105)	98.2	0.3	219	35785	44.3	3641863	d Bacteria;p Bacteroidota;c Bacteroidia;o Flavobacteriales;f UA16;g ;s unknown
13_P102_P105_crocinitomicaceae	Coassembly (102 + 105)	97.8	2.8	115	59205	41.1	3641863	d Bacteria;p Bacteroidota;c Bacteroidia;o Flavobacteriales;f Crocinitomicaceae;g UBA4466;s unknown
14_P102_P105_flavobacteriaceae	Coassembly (102 + 105)	97.9	1.7	132	29412	31.1	2365969	d Bacteria;p Bacteroidota;c Bacteroidia;o Flavobacteriales;f Flavobacteriaceae;g SCGC-AAA160-P02;s unknown
15_P102_P105_maricaulaceae	Coassembly (102 + 105)	99.5	0.0	8	627500	62.7	3372584	d Bacteria;p Proteobacteria;c Alphaproteobacteria;o Caulobacterales;f Maricaulaceae;g Maricaulis;s Maricaulis maris
16_P105_saccharospirillaceae	Single assembly (105)	99.0	1.8	45	122390	46.6	3878442	d Bacteria;p Proteobacteria;c Gammaproteobacteria;o Pseudomonadales;f Saccharospirillaceae;g Thalassolituus;s Thalassolituus oleivorans

Supplementary Information Table 2: Metagenome assembled genome (MAG) information and statistics. Quality metrics for MAGs were determined using CheckM[2] (Version 1.0.7; default parameters). The taxonomic identity of each MAG was determined using GTDB-Tk[3] (Version 1.0.2) against The Genome Taxonomy Database[4] (Version r89).

Metagenome	Contig#_ORF#	Assigned KO ID	KO Threshold	Hmmsearch Score	E-value	KO definition
14_P102_P105_flavobacteriaceae	c_000000000444_9	K00499	313.87	319.3	2.00E-99	choline monooxygenase
15_P102_P105_maricaulaceae	c_000000000002_229	K00499	313.87	367	6.30E-114	choline monooxygenase
4_P61_P62_pseudohongiellaceae	c_000000000337_16	K00499	313.87	173.9	2.60E-55	choline monooxygenase
6_P61_P62_rhodobacteraceae	c_000000000007_301	K00479	305.97	552.6	2.30E-170	glycine betaine catabolism
6_P61_P62_rhodobacteraceae	c_000000000010_117	K00479	305.97	269.4	2.40E-84	glycine betaine catabolism
6_P61_P62_rhodobacteraceae	c_000000000010_118	K00479	305.97	269.8	1.80E-84	glycine betaine catabolism

Supplementary Information Table 3: Further analysis of hits for *rhdA* (ring-hydroxylating dioxygenase subunit A) within metagenomes. KofamScan against the Kegg database hmms was performed for each hit to PF00848 (ring-hydroxylating dioxygenase subunit A). Resulting Kegg Orthology (KO) IDs and definitions indicate *rhdA* hits are for genes related to amino acid metabolism and catabolism. This is explained by the seed alignment for PF00848 (*rhdA*) including dioxygenases for aromatic hydrocarbon catabolism as well as genes for cytochrome P450 oxidoreductase, glutathione S-transferase, choline monooxygenase, proline dehydrogenase, as well as fairly ambiguous genes containing Rieske-type iron-sulfur clusters.

Analysis/ experiment	Pentadecane Concentration	Pentadecane Production	Particle Pentadecane Content	Nutrients, cell counts	Pentane Respiration	Pentadecane Respiration and Metagenomics
Station 1 (40°25.273' N, 68°11.855'W)	--	--	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	--
Station 2 (40°9.14' N, 68°19.889'W)	--	--	150 m	100%, 30%, 10%, 3%, 1%, 0.3% PAR	1000 m	---
Station 3 (36°52.17' N, 71°23.94'W)	30%, 10%, 1% PAR	30%, 10% PAR	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	500 m
Station 4 (33°59.929' N, 69°58.014'W)	30%, 10%, 1% PAR and 25 m, 45 m, 60 m, 75m, 90 m, 105 m, 125 m, 145 m	30%, 10%, 1% PAR	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR	1000 m	--
Station 5 (31°41.34' N, 70°46.106'W)	30%, 10%, 1% PAR	30%, 10%, 1% PAR	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	--
Station 6 (29°14.038' N, 69°58.497'W)	100%, 30%, 10%, 3%, 1%, 0.3% PAR and DCM	--	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR and DCM	--	500 m
Station 7 (29°2.682' N, 66°3.110'W)	30%, 10%, 1% PAR	10%, 1% PAR	150 m	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	--
Station 8 *BATS location* (32°10.815' N, 64°09.807'W)	30%, 10%, 1% PAR and 10 m, 20 m, 30 m, 40 m, 50 m, 65 m, 75 m, 85 m, 95 m, 130 m, 150 m, 200 m, 700 m	30%, 10%, 1% PAR	150 m	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	--
Station 9 (38°30.74' N, 68°0.890'W)	100%, 30%, 10%, 3%, 1%, 0.3% PAR and DCM	--	--	100%, 30%, 10%, 3%, 1%, 0.3% PAR	--	--
Station 10 *Gulf of Mexico* (27°30.41'N, 87°12.41'W)	--	--	--	1000 m	1000 m	--
Station 11 *Gulf of Mexico* (27°15.00'N, 89°05.05'W)	--	--	--	1000 m	1000 m	--
Station 12 *Gulf of Mexico* (27°11.60'N, 90°41.75'W)	--	--	--	1000 m	1000 m	--
Station 13 *Gulf of Mexico* (27°38.40'N, 90°54.98'W)	--	--	--	1000 m	1000 m	--

Supplementary Information Table 4: Overview of oceanographic experiments. This table indicates each type of analysis conducted in this study at each of the 13 stations and the depths/ light levels at each station where seawater was collected. Pentadecane concentration analysis was conducted at n = 2 replication for all depths described as "meters" for depth profiles and station 9 and 6, the remainder were conducted in n = 3 biologically independent replication. Pentadecane production measurements were generally conducted in n = 3 replication except for station 3 (n = 2 for 30% PAR, n = 1 for 10% PAR). Pentadecane content of particles were quantified at n = 1 replication. Cell counts and nutrients were conducted at n = 1 replication. All respiration experiments were conducted in n=6 and metagenomics was used at n=2 replication. See Extended Data Figure 8b for information regarding Tara Oceans alkane-1-monooxygenase analysis. Key to abbreviations and symbols: -- indicates no measurements were taken; m stands for meters and

indicates a sampling depth; % PAR indicates the light penetration depth at which samples were collected and the light attenuation level used for the corresponding incubation experiments; DCM refers to the sampling targeted to the deep chlorophyll maximum.

Supplementary Note

Chemical extraction and data quality check

We performed 441 chemical extractions of hydrocarbons from 441 individual filters. Extraction efficiency, defined as the percent recovery of the internal standard dodecahydrotriphenylene (DDTP), averaged 83.9% with a standard deviation of 16.7% (Extended Data Fig. 1a). Overall, replicates (as triplicates for production experiments or duplicates for depth profiles) showed high precision (average standard deviation = 3 ng L^{-1} pentadecane), with only 6 replicates having a standard deviation > 11 ng L^{-1} pentadecane (Extended Data Fig. 1b). Larger standard deviations within replicates were associated with larger cyanobacteria cell abundances (Extended Data Fig. 1d) and thus pentadecane concentrations. Analysis of these replicates revealed that two of them contained an outlier well explained by human error when compared to five other environmentally identical samples and were thus excluded from further use. The other two data were associated with very large pentadecane concentration (~100 ng/L) and thus had lower standard error; they showed no obvious human error and we chose to keep these data. All other data showed no clear issues with extraction or sampling procedures.

Pentadecane concentration

In this work pentadecane concentration is expressed as mass per volume of filtered seawater. However, it is important to note that pentadecane (and heptadecane) is interpreted to

reside primarily in the membranes of cells in the particulate phase[5]. Thus, these units do not represent a truly dissolved chemical compound.

Pentadecane production in the Gulf Stream

The sampling station located in the Gulf Stream exhibited high production at 10% PAR (50 ng nC_{15} L^{-1} d^{-1}), which aligned with the DCM at ~50 m (Fig. 1). In contrast, the "true" oligotrophic stations exhibited a 1% PAR depth usually aligning close to the DCM (within ~15 m). Furthermore, a large *Prochlorococcus* population was seen with higher nitrite concentrations at 10% PAR in the Gulf Stream compared to the "true" oligotrophic stations. Station 7 exhibited lower production at 1% PAR (10 ng nC_{15} L^{-1} d^{-1}) compared to other oligotrophic stations (Fig. 3a) likely due to a comparatively smaller *Prochlorococcus* population at this depth and extremely low nutrients. These anomalies are well reconciled in the multiple linear model predicting cell-specific production rates from dissolved nitrite and cellular pentadecane content (Fig. 2e).

Continental shelf waters

Station 1 was the sole station located in eutrophic waters on the continental shelf. These waters are dominated by eukaryotic phytoplankton and thus were not the focus of our study. Furthermore, we did not utilize a 200 μm mesh to catch large zooplankton at this station and as a result observed variable amounts of zooplankton in our 2L samples. Additionally, we observed a chromatographic coelution in these samples with both heptadecane and our internal standard, DDTP. We found that heptadecane was present at higher concentrations than in nutrient poor waters and was always higher in concentration than pentadecane at this station (although with a similar ~3:1 ratio), consistent with eukaryotic-derived octadecanoic acid (stearic acid) as the precursor. Additionally, heptadecane exhibited more variable concentrations between replicates

which might be related to the presence of large zooplankton or other forms of heterogeneity in these waters. For these reasons we excluded this data from this study, and refer to the results in qualitative terms.

The use and rationale for steady state calculation of pentadecane production using ^{13}C

Considering that 30% and 10% PAR waters were at a steady state with respect to pentadecane concentration, we used a modified primary production calculation using ^{13}C enrichment from López-Sandoval et al.[6], to calculate the production of pentadecane. The concentration of pentadecane for the 1% PAR incubation increased over the 30-hour incubation for most oligotrophic stations (Extended Data Fig. 2), violating an assumption outlined by López-Sandoval et al (2018). We thus chose to compare two approaches: calculation of production via concentration data only and calculation using ^{13}C from López-Sandoval et al., 2018. Ultimately, we chose to use the isotope-predicted production rates of pentadecane because loss processes were clearly evident from the comparison of the two approaches.

Diel patterns for pentadecane, cells and fluorescence

Pentadecane concentrations were consistent over the diel cycle for shallower depths but not for the deep photic zone (DCM, 1% PAR and 3% PAR), which varied between 8-30%, with the DCM displaying the greatest change (Extended Data Fig. 4). Density variations were minor for waters sampled at the DCM, 1% PAR and 3% PAR depths (Extended Data Fig. 5) supporting the interpretation that observed variations reflect biological process rather than sampling bias or physical processes.

Like pentadecane, the abundance of *Prochlorococcus* remained consistent in the upper portion of the photic zone throughout the diel cycle whereas it varied substantially (~50%) in the

lower photic zone (DCM, 1% and 3% PAR depths) (Extended Data Fig. 4b). *Prochlorococcus* abundance in the lower photic zone was observed to decrease in the daylight hours by ~half, with replenishment beginning at dusk and continuing through the night, to meet the original concentration at dawn (Extended Data Fig. 4b). This behavior is reflective of previously reported doubling patterns of *Prochlorococcus* in both the laboratory and in the ocean[7,8], with cell growth during the day and the peak of cell division occurring near dusk. *Synechococcus* abundance did not follow any discernable pattern over the diel cycle (Extended Data Fig. 4c) and was approximately an order of magnitude lower in abundance than *Prochlorococcus*, thus we interpret pentadecane dynamics to stem primarily from *Prochlorococcus*.

Combining pentadecane concentrations with *Prochlorococcus* abundance patterns enables an assessment of cell-specific pentadecane variability over a diel cycle at three depths (DCM, 1% PAR, 3% PAR). Accumulation of pentadecane preceded cell division, consistent with diurnal growth preceding nocturnal division. This pattern is also in alignment with previous reports from laboratory knockout experiments of *Synechocystis* (a freshwater cyanobacterium) indicating that hydrocarbons promote membrane flexibility and optimal cell growth and division[5]. Thus, it appears that the diel changes in the pentadecane per *Prochlorococcus* cell measured here are reflective of *Prochlorococcus'* cell physiology relative to day-night cycles of growth and division. Since *Prochlorococcus* abundance is a balance between cell death and cell division, cell-specific pentadecane production rates and average cell division rates may serve as reasonable scaling factors to calculate hydrocarbon production in the ocean. The cell-specific content of pentadecane was notably higher in waters at the DCM at all times in the diel cycle, compared to other depths (Extended Data Fig. 4e). This observation further highlights our finding that pentadecane abundance is proportional to fluorescence, and may shed light on utilization of hydrocarbons for

photo-acclimation by cyanobacteria in low-light environments. Specifically, these results are consistent with a model in which increased membrane stacking serves as a low-light adaptation, housing more chlorophyll and requiring more alkane to minimize curvature stress.

We also find that 1% and 3% PAR waters exhibit an increase in pentadecane/fluorescence during the day (Extended Data Fig. 4f), with a decrease at night. This observation could be interpreted as an increase in internal membranes (scaffolding) and the need for tight membrane curvature preceding production of chlorophyll and division, however, an opposing trend was observed for the DCM and the topic warrants further investigation. Regardless, it is clear that pentadecane concentrations are stable in the upper photic zone and with lower concentrations as compared to the lower photic zone in which there is more rapid production and utilization of pentadecane by cyanobacteria, particularly *Prochlorococcus*. These results further bolster our finding that the DCM is a highly dynamic focal point for biogeochemical cycling of pentadecane.

Reconciling discrepancies between Biogeochemical Estimate Methods 1 and 2

The water column integrated approach is representative of pentadecane stock in the open-ocean oligotrophic gyres insomuch as the locations (in the North Atlantic subtropical gyre), season and local conditions are scalable, yielding values smaller than the cellular specific calculated stock (by a factor of 0.4, Supplementary Information Table 1). The reason for this discrepancy is uncertain though partially attributable to the occurrence of *Prochlorococcus* and *Synechococcus* in regions outside the open-ocean oligotrophic gyres, particularly for *Synechococcus* which is found in both coastal oligotrophic waters and more eutrophic waters. Population differences between ocean basins (Extended Data Fig. 10) may also be an important contributor. We propose that our method using water column integration is a reasonable representation for the North

158 Atlantic subtropical gyre but our average estimate using modeled cell-specific concentrations may
159 be more accurate for the global stock for oligotrophic ocean regions.

160 **Microbial productivity from pentadecane**

161 Based on the assumption of steady state, we estimated the magnitude for the production
162 rate of obligate alkane-degrading bacteria or archaea using cyanobacterial pentadecane as sole
163 substrate in the oligotrophic ocean. Assuming a carbon conversion efficiency range (pentadecane
164 to biomass) of 5-50% and a carbon mass of hydrocarbon degrading microbes of 120 fg C $cell^{-1}$ ^{1}
165 pentadecane in the lower photic zone (1% PAR) would support microbial production of the order
166 of ~10-100 cells ml^{-1} d^{-1} with the upper photic zone (30% PAR) supporting ~2-20 cells ml^{-1} d^{-1}.
167 The size of the supported community further depends on cellular turnover time, and an assumed
168 turnover rate of 0.1 day^{-1} equates to a steady state population of ~10^2-10^3 cells ml^{-1} in the lower
169 photic zone (1% PAR) and ~20-200 cells ml^{-1} in the upper photic zone (30 % PAR) of the
170 oligotrophic ocean (Supplementary Information Table 1). These results underscore the depth
171 dependency of cyanobacterial pentadecane production, and the potential for similar structuring for
172 the microbial community of alkane degraders. Furthermore, we expect a secondary structuring of
173 alkane degradation based on phase-state of the alkanes, with particles showing more consumption
174 than surrounding waters, as we interpret in our incubation experiments and sediment trap data.

175 **Microbial community within pentadecane incubations**

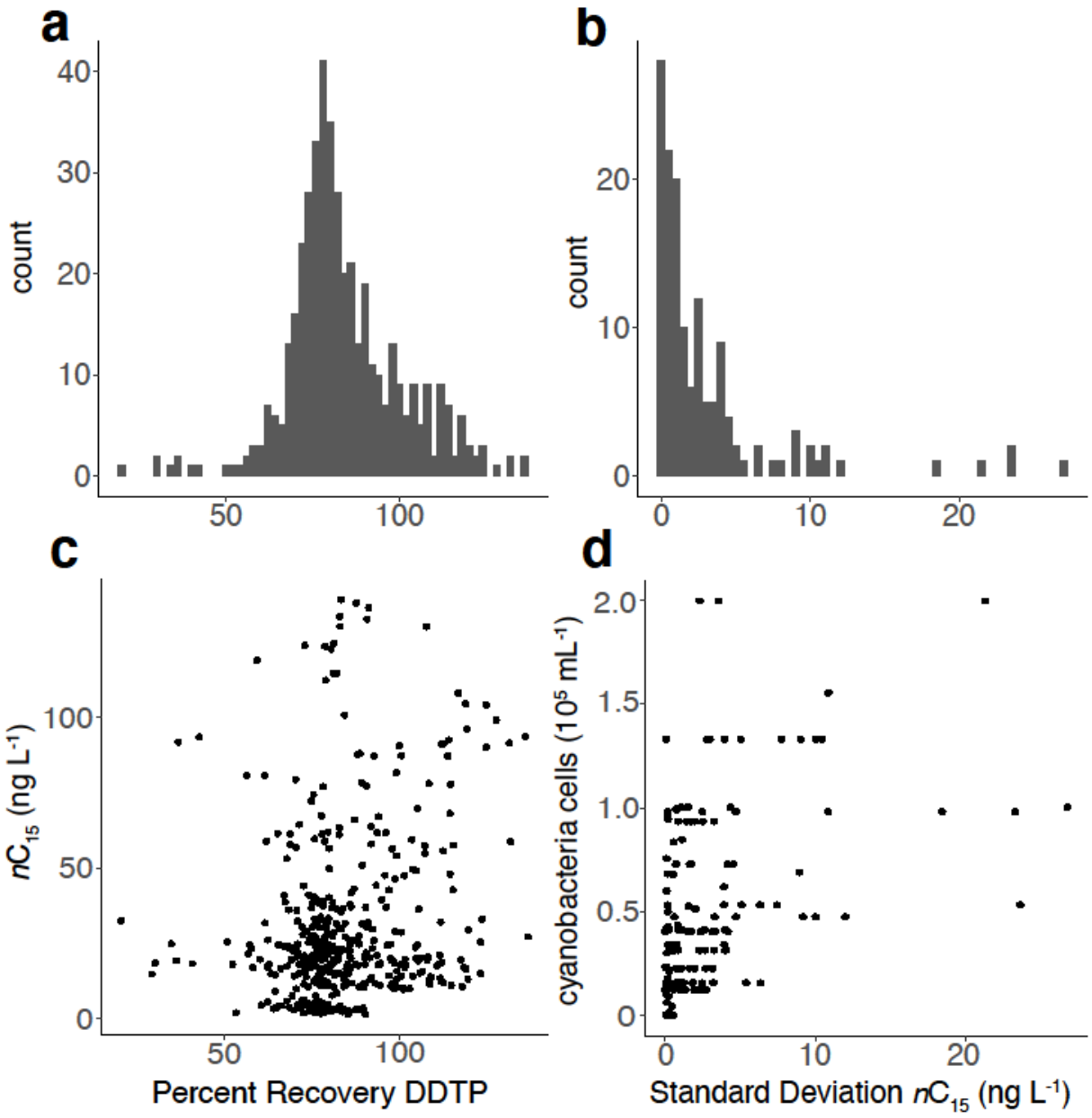
176 The pentadecane + particle treatments at both sites elicited response by a greater number
177 of taxa compared to pentadecane only treatments. This response was quantified using Shannon
178 diversity indices whereby particle + pentadecane enrichments had a higher Shannon diversity
179 index compared to the pentadecane only treatments (two tailed t-test: Station 3, $p < 0.023$; Station

6, $p < 0.017$) (Extended Data Figure 6). Presumably, the particles provided a diverse range of substrates beyond the added pentadecane and/or more taxa that could consume pentadecane. Nonetheless, several taxa that contain opportunistic hydrocarbon degraders were observed to bloom in these treatments including *Litoribacillus* (*Saccharospirillaceae*)*, Pseudophaeobacter* (*Rhodobacteraceae*)*,* and *Aurantivirga* (*Flavobacteriaceae*) at station 3 and *Thalassolituus* (*Saccharospirillaceae*) and *Olleya* (*Flavobacteriaceae*) at station 6. From this data, we postulate a possible association of alkane specialists with the hydrophobic phase and alkane generalists with particulate matter.

188 **Supplementary References**

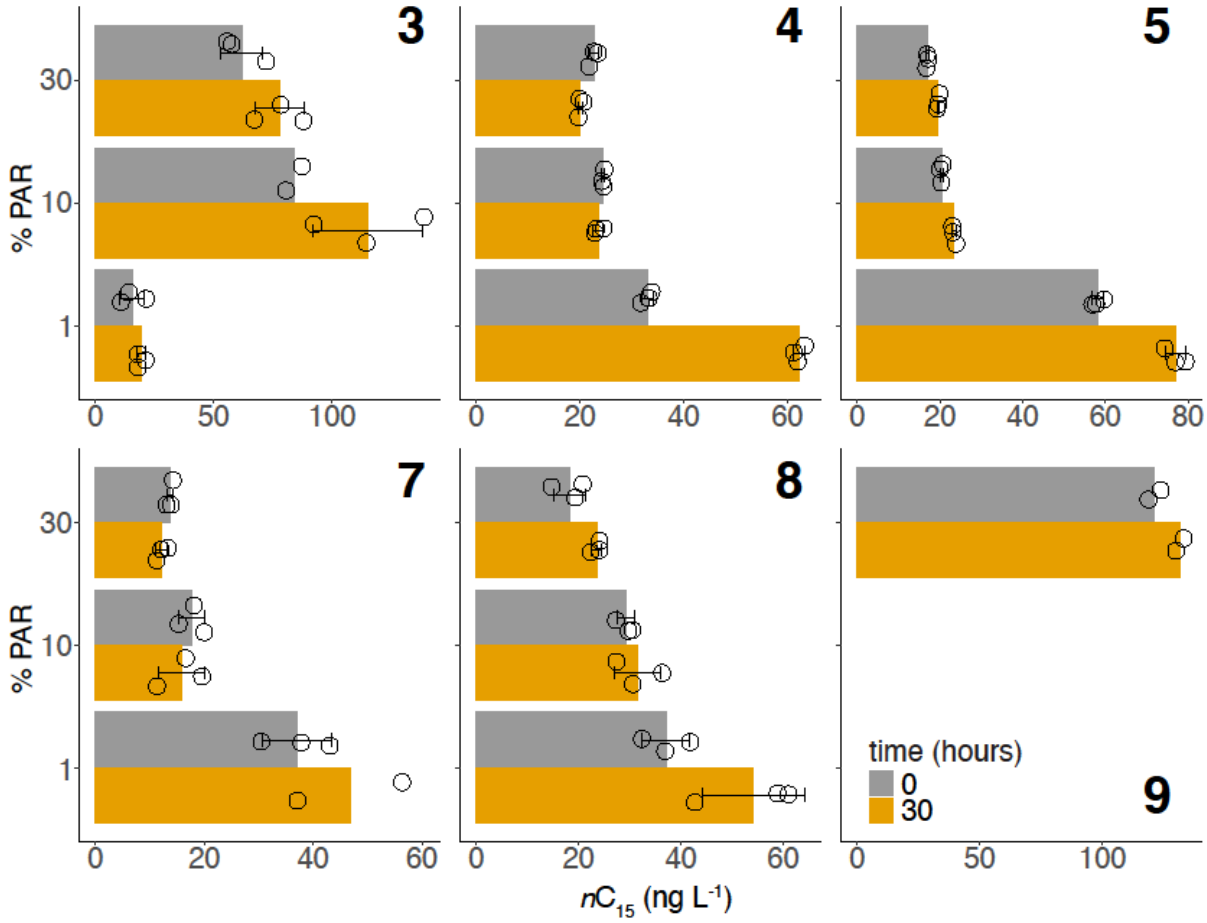
189 1. Valentine, D. L. *et al.* Dynamic autoinoculation and the microbial ecology of a deep water
190 hydrocarbon irruption. *Proceedings of the National Academy of Sciences of the United*
191 *States of America* **109**, 20286–20291 (2012).
192 2. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM:
193 Assessing the quality of microbial genomes recovered from isolates, single cells, and
194 metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
195 3. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to
196 classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927
197 (2019).
198 4. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat.*
199 *Biotechnol.* 1–8 (2020). doi:10.1038/s41587-020-0501-8
200 5. Lea-Smith, D. J. *et al.* Hydrocarbons are essential for optimal cell size, division, and growth
201 of Cyanobacteria. *Plant Physiol.* **172**, 1928–1940 (2016).
202 6. López-Sandoval, D. C., Delgado-Huertas, A. & Agustí, S. The 13 C method as a robust
203 alternative to 14 C-based measurements of primary productivity in the Mediterranean Sea.
204 *J. Plankton Res.* **40**, 544–554 (2018).
205 7. Zinser, E. R. *et al.* Choreography of the transcriptome, photophysiology, and cell cycle of
206 a minimal photoautotroph, Prochlorococcus. *PLoS One* **4**, 5135 (2009).
207 8. Ribalet, F. *et al.* Light-driven synchrony of Prochlorococcus growth and mortality in the
208 subtropical Pacific gyre. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8008–8012 (2015).
209
210

211 Extended Data Figure 1



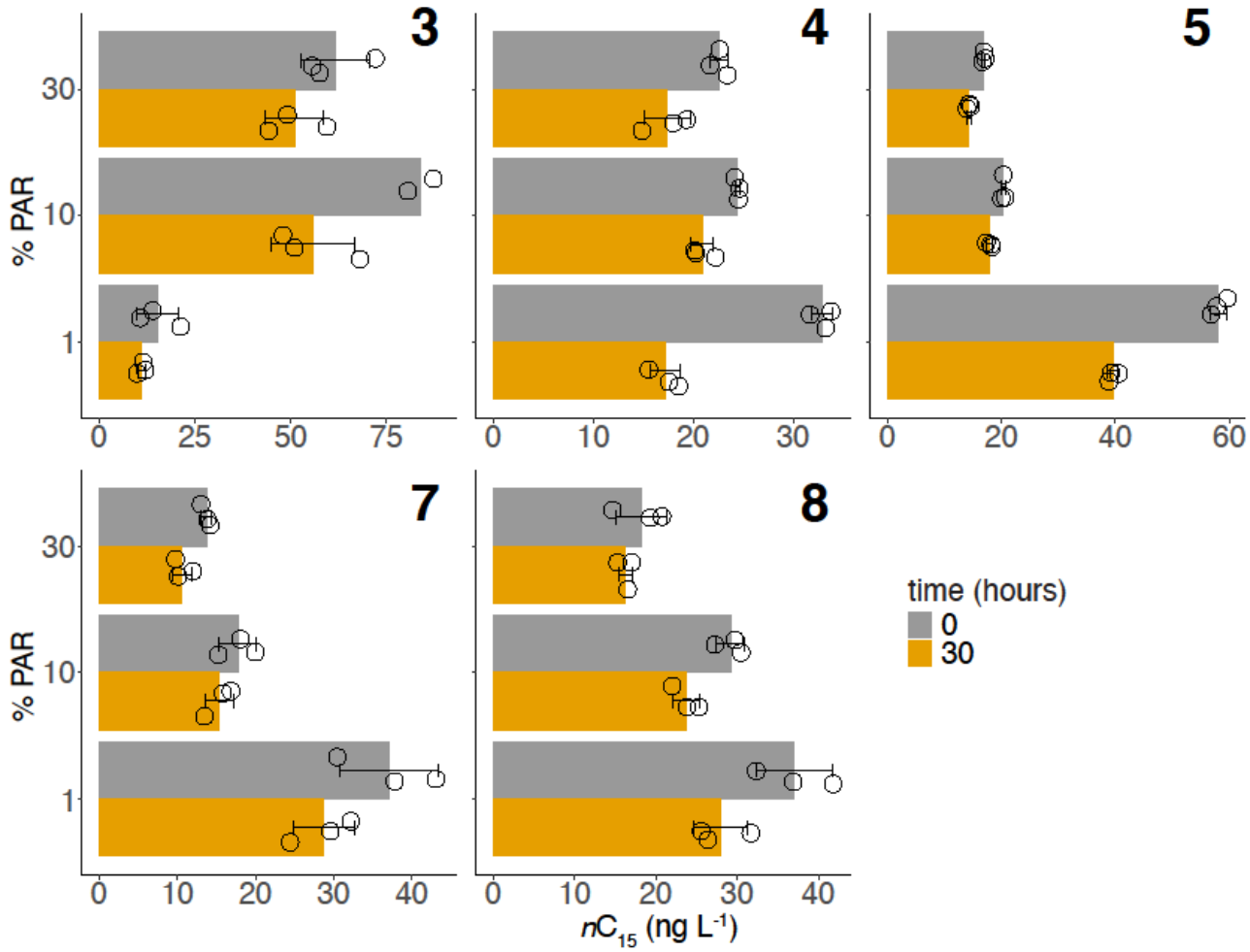
212

213 Extended Data Figure 2

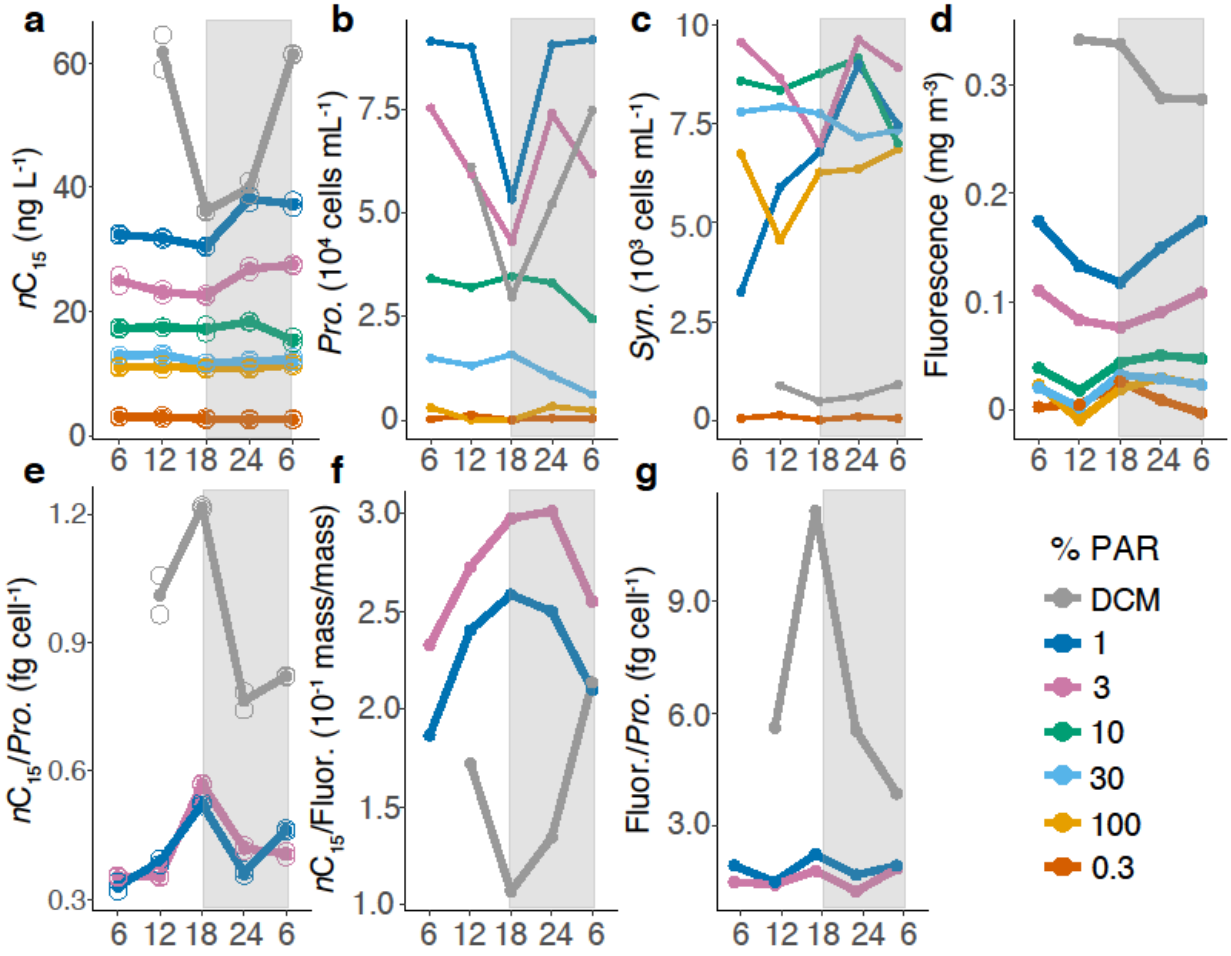


214

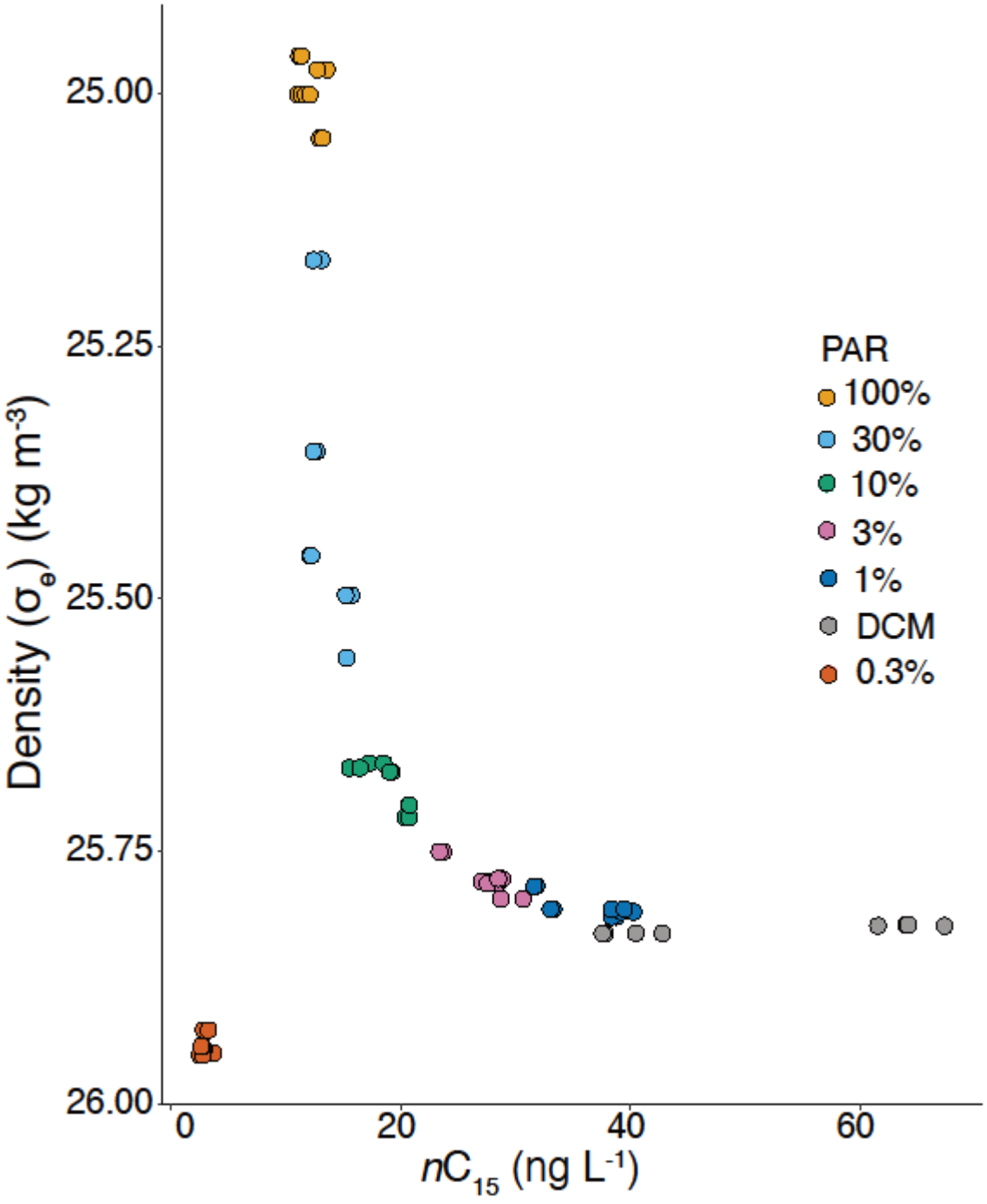
Extended Data Figure 3



Extended Data Figure 4

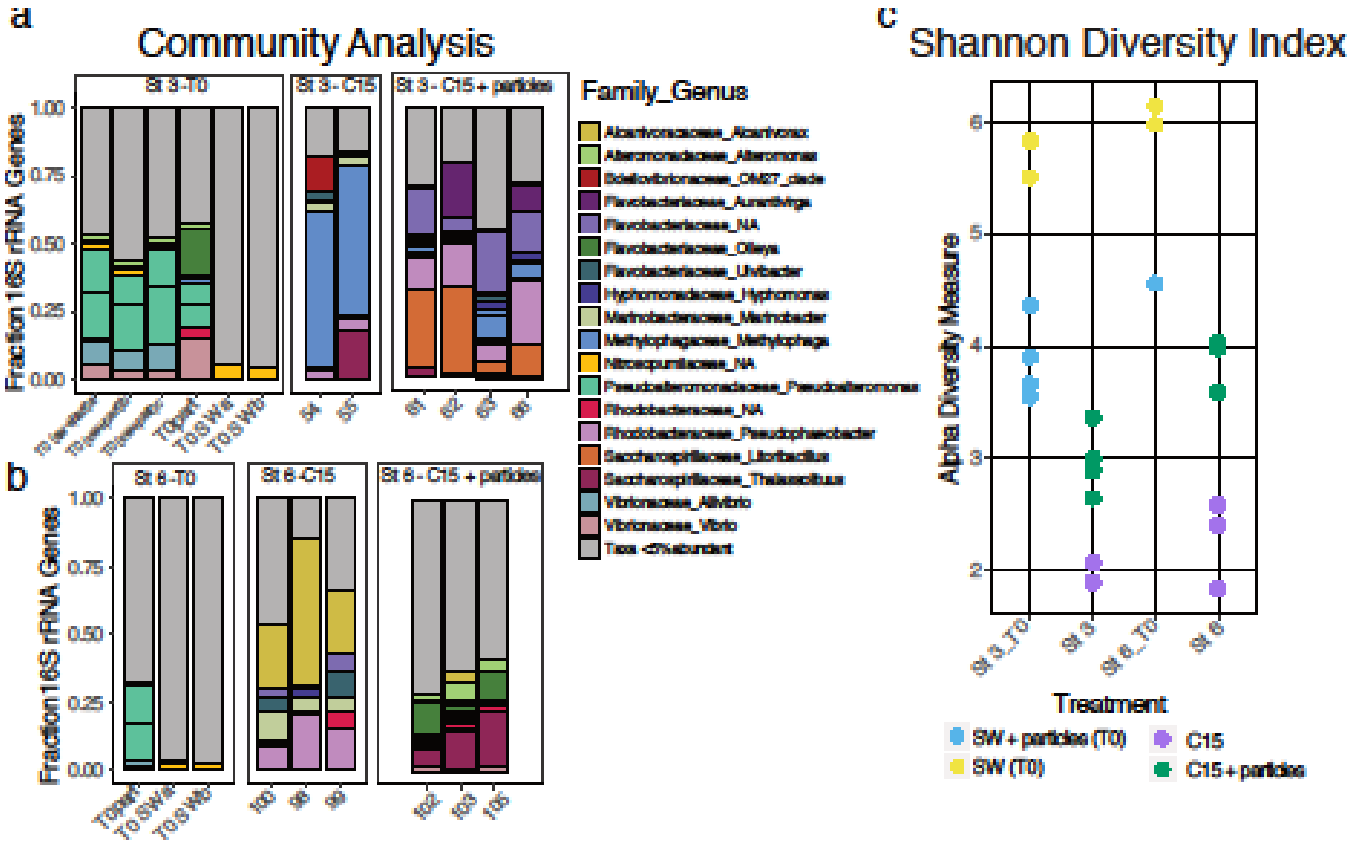


219 Extended Data Figure 5



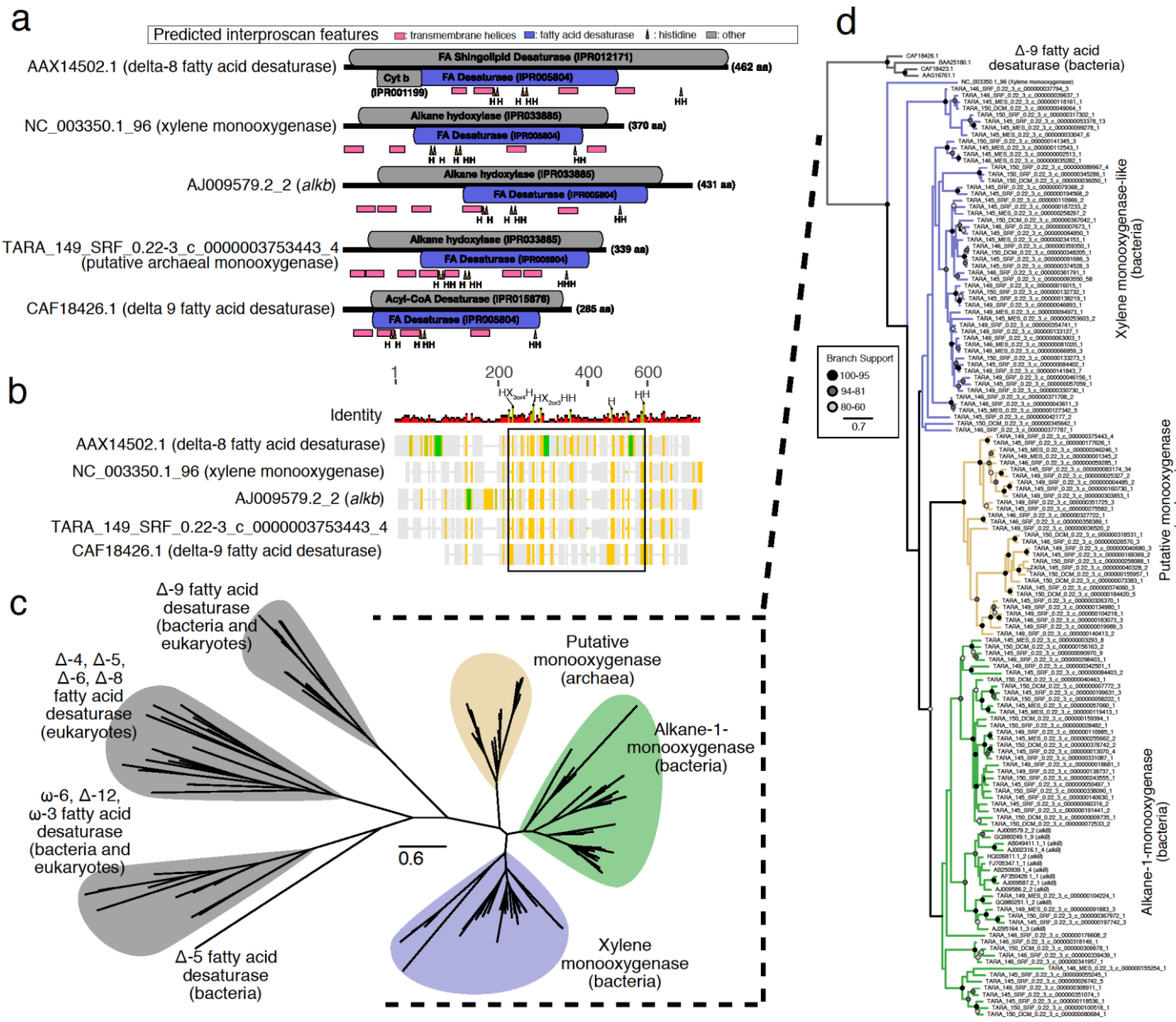
220

221 Extended Data Figure 6



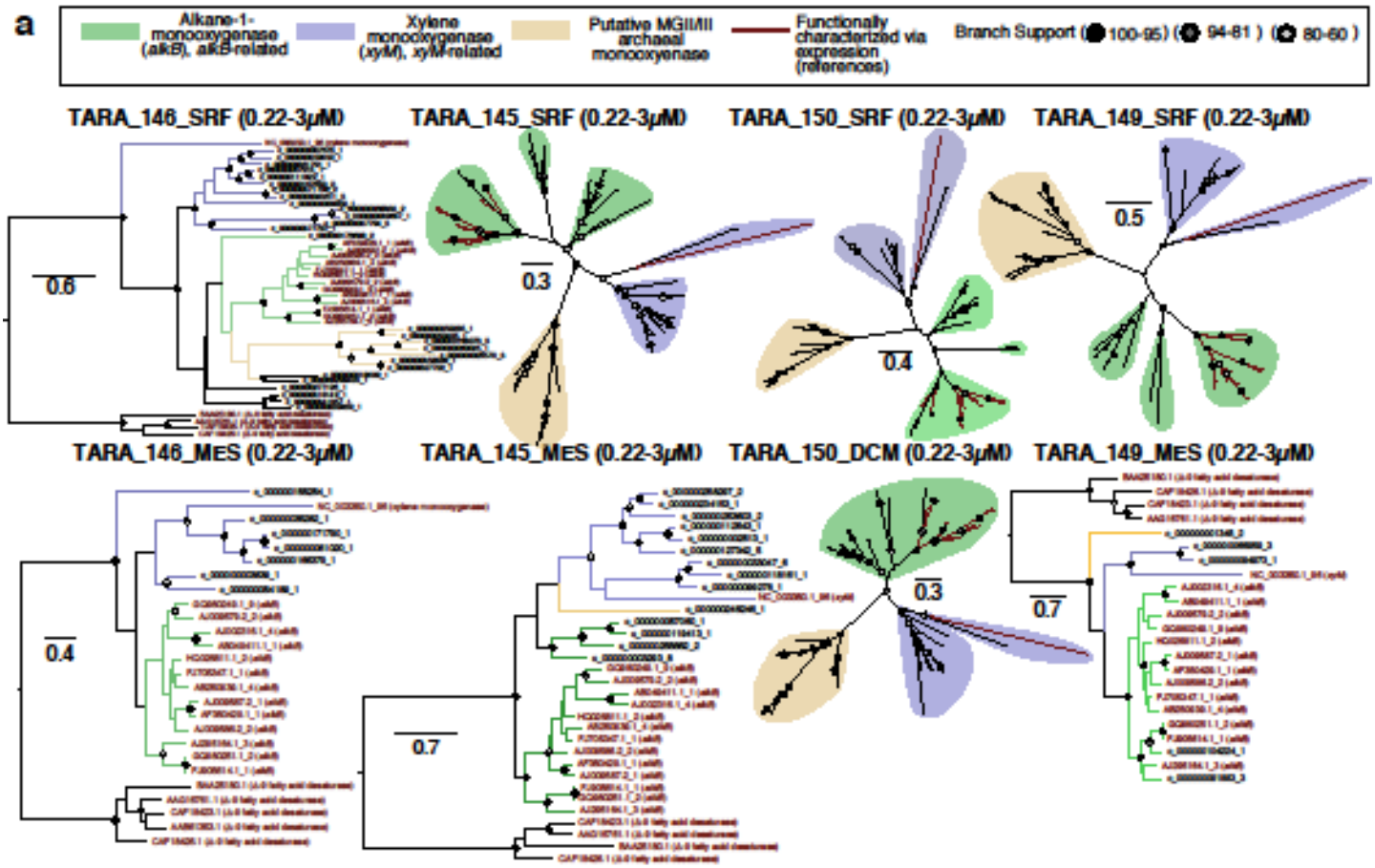
222

223 Extended Data Figure 7



224

Extended Data Figure 8



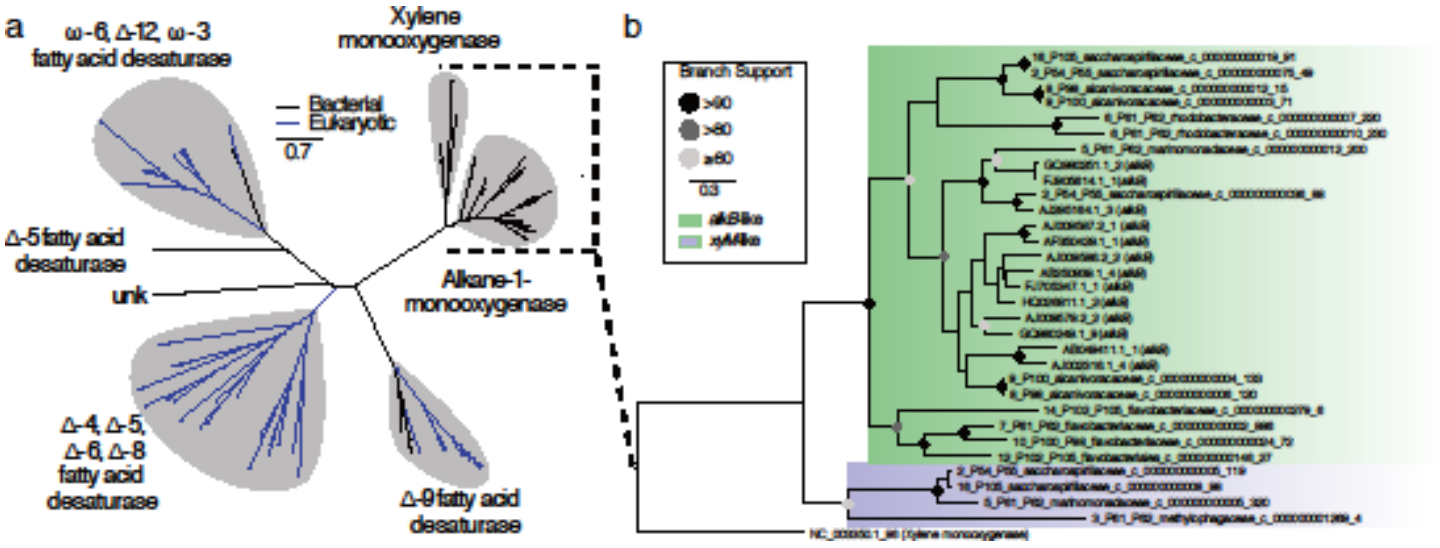
b

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	-71.3098	5	27.6	7
				59.7	19
				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	-49.9161	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	35.6427	-57.1526	40	40.7	14
				26.7	10
				32.6	13
TARA_150_SRF_0.22-3	35.9346	-57.3032	5	35.5	9
				30.0	9
				34.5	11

Clade Type
Archaea
Xylene
AlkB

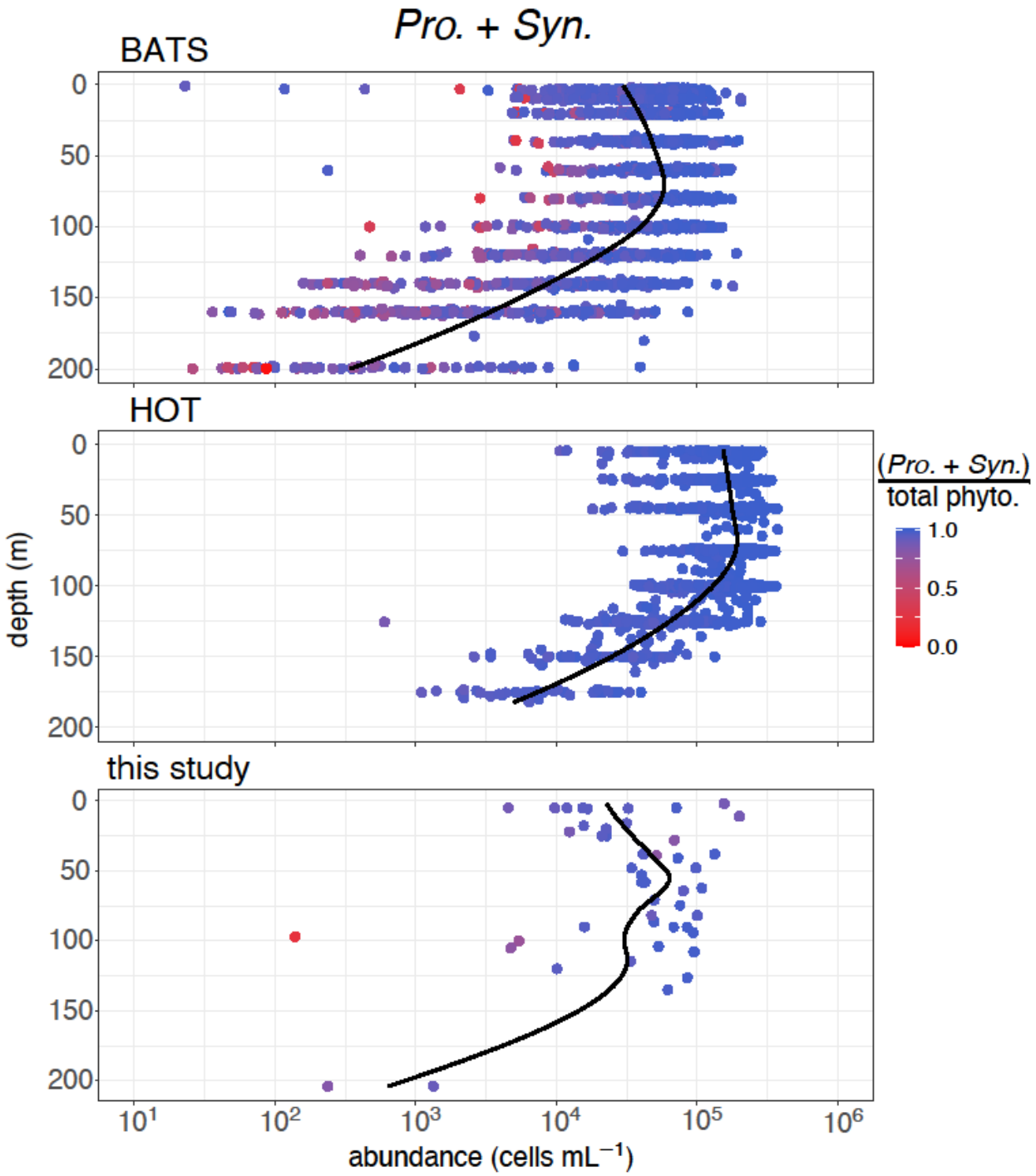
226

227 Extended Data Figure 9



228

229 Extended Data Figure 10



230

Extended Data Figure Captions

Extended Data Figure 1. 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 Figure 2. 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 Figure 3. 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 Figure 4. 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 Figure 5. 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 Figure 6. **a-b** Microbial community composition within pentadecane incubations informed via the V4 region of the 16S 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 Figure 7. 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 Online Methods). Each column of alignment figure represents a sliding window of 5bp 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#).

Extended Data Figure 8. **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 Figure 9. **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 Figure 3h.

Extended Data Figure 10. 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

315 from Hawaii Ocean Time-series HOT-DOGS application; University of Hawai'i at Mānoa,
316 National Science Foundation Award #1756517), and this study (at bottom). Data points are colored
317 on a gradient by the proportional contribution to the phytoplankton community by
318 *Prochlorococcus* and *Synechococcus* (total phytoplankton community is calculated as *Pro*. + *Syn*.
319 + pico- + nano-Eukaryotes for BATS and this study, and *Pro*. + *Syn*. + pico-Eukaryotes for HOT).
320 BATS and HOT data are each from a single station measured nearly monthly for ~20 years whereas
321 measurements from this study incorporate spatial variability (see Fig. 1) with minimal temporal
322 variability (all measurements taken in May 2017). The proportional contribution of
323 *Prochlorococcus* and *Synechococcus* is >90 % of the phytoplankton community at BATS 84% of
324 the time. At HOT, *Pro*. + *Syn*. is > 90 % of phytoplankton community ~100% of the time. For this
325 study, *Pro.* + *Syn*. is >90% of the phytoplankton community (*Pro*. + *Syn*. + pico + nano-
326 Eukaryotes) for 80% of the measurements, with most cases of lower proportional prokaryote
327 abundance due to an anomalous nutrient pulse observed at station 9 (a *Synechococcus* bloom) or
328 at low absolute abundance of *Pro*. + *Syn*.