

Metabolic and spatio-taxonomic response of uncultivated seafloor bacteria following the Deepwater Horizon oil spill

Authors

Handley KM^{a,b,c,§}, Piceno YM^d, Hu P^d, Tom LM^d, Mason OU^e, Andersen GL^d, Jansson JK^f, Gilbert JA^{a,b,g,1}

Author Affiliation

^aSchool of Biological Sciences, University of Auckland, Auckland 1010, New Zealand

^bDepartment of Ecology and Evolution, The University of Chicago, Chicago, IL 60637, USA

^cInstitute for Genomic and Systems Biology, Argonne National Laboratory, IL 60439, USA

^dEarth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^eEarth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL 32304, USA

^fEarth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA

^gThe Microbiome Center, Department of Surgery, The University of Chicago, Chicago, IL 60637, USA

¹Correspondence should be addressed to Jack A. Gilbert, Ph.D. Department of Surgery, University of Chicago, 5842 South Maryland Avenue, Chicago, IL, 60637, USA; email: gilbertjack@uchicago.edu; and Kim M. Handley, School of Biological Sciences, University of Auckland, Auckland, New Zealand, email: kim.handley@auckland.ac.nz

1 **Abstract**

2 The release of 700 million liters of oil into the Gulf of Mexico over a few months in 2010
3 produced dramatic changes in the microbial ecology of the water and sediment. Previous
4 studies have examined the phylogeny and function of these changes, but until now a
5 fundamental examination of the extant hydrocarbon metabolisms that supported these
6 changes had not been performed. Here, we reconstructed the genomes of 57 widespread
7 uncultivated bacteria from post spill sediments, and recovered their gene expression
8 pattern across the seafloor. These genomes comprised a common collection of bacteria
9 that were highly enriched in heavily affected sediments around the wellhead. While rare
10 in distal sediments, some members were still detectable at sites up to 60 km away. Many
11 of these genomes exhibited phylogenetic clustering indicative of common trait selection
12 by the environment, and within half we identified 264 genes associated with hydrocarbon
13 degradation. Observed alkane degradation ability was near ubiquitous among candidate
14 hydrocarbon degraders, while just 3 harbored elaborate gene inventories for the
15 degradation of alkanes and (poly)aromatic hydrocarbons. Differential gene expression
16 profiles revealed a spill-promoted microbial sulfur cycle alongside gene up-regulation
17 associated with polyaromatic hydrocarbon degradation. Gene expression associated with
18 alkane degradation was widespread, although active alkane degrader identities changed
19 along the pollution gradient. The resulting analysis suggests a broad metabolic capacity
20 to respond to oil exists across a large array of usually rare bacteria.

21 Marine oil spills are frequent occurrences that can have a severe impact on environmental
22 health and dependent economies¹. In US marine environments alone, hundreds of oil
23 spills occur annually, releasing millions of liters of oil per year on average². While spills
24 have typically occurred at shallow water depths, an expansion of drilling into the deep-
25 sea led to the 2010 Deepwater Horizon (DWH) accident in the Gulf of Mexico³. DWH
26 released over 700 million liters of oil from the Macondo MC252 wellhead at 1500 m
27 depth⁴. This yielded a vast sea surface oil slick⁵, and an expansive plume of hydrocarbons
28 at a water column depth of ~1100 m⁶. The leak also polluted deep-sea sediments up to
29 tens of kilometers distance, due to direct contamination and flocculent fallout from
30 plumes⁷. Pollution of seafloor sediments persisted at least 3 months post spill⁸, likely
31 supported by ongoing inputs from sinking hydrocarbon-bearing particles⁹. Post spill
32 contamination of sediment was greater than in the water column and was greatest within
33 3 km of MC252, where total polyaromatic hydrocarbon (PAH) concentrations remained
34 above the Environmental Protection Agency's Aquatic Life benchmark^{8,10}.

35 Natural bacterial communities are important agents for breaking down the complex
36 mixtures of hydrocarbons in leaked oil^{11,12}. Extensive degradation of petroleum
37 hydrocarbons released during the DWH spill has been largely attributed to microbial
38 activity¹³. Gammaproteobacteria – some clearly related to psychrotolerant or
39 psychrophilic bacteria – dominated the deep-sea response^{10,14,15}, contrasting with higher
40 alphaproteobacterial ratios in shallow environments¹⁶. Although the buoyant plume
41 received far greater attention than the underlying seafloor sediments¹³, post-spill
42 sediments were observed to share some key taxa with the plume; a bacterium associated
43 with the hydrocarbonoclastic genus *Colwellia* and an abundant undescribed
44 gammaproteobacterium¹⁰. Genes or transcripts associated with hydrocarbon degradation
45 were enriched in both the plume and polluted sediments^{10,17,18}, and some were linked to
46 single cell genomes of two plume alkane-degrading *Colwellia* and Oceanospirillales
47 bacteria^{18,19}. However, the communitywide and organism-specific metabolic response to
48 the spill has not been directly explored leaving much of the hydrocarbon degrading
49 potential of numerous uncultivated oil-responsive bacteria enigmatic, particularly on the
50 seafloor.

51 Here we used 13 metagenomes¹⁰ and 10 metatranscriptomes to link communitywide
52 hydrocarbon degradation strategies with microbial taxonomy in the top one centimeter of
53 deep-sea sediment, collected 3 months after the DWH wellhead was capped. We
54 reconstructed bacterial genomes from the metagenomes, and mapped metatranscriptomes
55 to these genomes to determine the activity of oil-responsive functional pathways across a
56 hydrocarbon concentration gradient left by the spill. Studied communities included those
57 from seven highly polluted 'near-well' sites around MC252 (0.3 to 2.7 km away), and six
58 'distal' sites that were distributed along a linear transect (10.1 to 59.5 km away)¹⁰, which
59 followed the prevailing southwesterly deep plume path⁶ (Fig. S1). Distal-most sites were
60 either un-impacted or minimally impacted, based on greatly diminishing plume
61 hydrocarbon concentrations beyond ~30 km^{7,20} and oil-proxy sediment hopane
62 concentrations beyond 40 km from the wellhead⁷. Results provide insights into the
63 genomic potential and in situ transcriptional activity of dozens of spill-responsive
64 bacteria, including more than 20 candidate hydrocarbon degraders.

65
66

67 **Results and Discussion**

68 To establish a site-specific genomic database for metatranscriptome mapping,
69 metagenomic sequences were co-assembled from three genomically representative (Fig.
70 1) and comparatively well-assembling sediment samples (Table S1) collected 0.5, 0.7 and
71 0.9 km from MC252. Contigs were binned into genomes aided by compositional
72 information²¹ and differential coverage²², which was obtained by mapping metagenomic
73 reads from all 13 sites to the co-assembly. This also enabled us to determine the
74 abundance of co-assembled genomes at each site. Most sequences (66% or 119 Mbp)
75 were classified into 51 bins comprising 57 genomes (Table S2, Dataset S1) associated
76 with the Gammaproteobacteria, Alphaproteobacteria, Deltaproteobacteria and
77 Bacteroidetes (Table S2). A further 4.4% of contigs contained virus-like genes that were
78 mainly associated with the Gammaproteobacteria. The relative abundance of the bacterial
79 genome bins ranged from 0.6 to 13.1% (1.6% on average), and consisted of partial to
80 near complete genomes estimated to be 7 to 97% complete (50% on average; Table S2).
81 Five bins contained 2 to 3 very closely related genomes with partial (insufficient)
82 coverage separation (Fig. S2). Metatranscriptomic sequences, derived from 10 of the
83 same samples as the metagenomes, were then mapped to the co-assembly, generating
84 genome-specific expression profiles up to 33.9 km from MC252 (Dataset S1). This also
85 enabled mRNA hits to be normalized to genome abundance per site to determine gene
86 up/down regulation.

87 Small subunit (SSU) rRNA genes and rRNA sequences from the metagenomes and
88 metatranscriptomes were reconstructed using EMIRGE²³, and co-clustered into
89 operational taxonomic units (OTUs), including rRNA sequences from an additional
90 unpaired metatranscriptome sample collected 1.3 km from MC252. Comparison of the
91 relative abundances of bacterial, archaeal and eukaryotic SSU rRNA gene sequences
92 demonstrated that bacteria near MC252 increased appreciably relative to archaea and
93 eukaryotes (Fig. S3). Gammaproteobacteria predominated near MC252¹⁰ and exhibited
94 increased species richness (Fig. S3), analogous to the predominance of
95 Gammaproteobacteria observed in the deep-sea plume¹⁴. While rRNA is not necessarily a
96 good indicator of microbial growth²⁴, we found rRNA gene and rRNA relative
97 abundances were generally well correlated for prokaryotes (but not eukaryotes),
98 particularly for bacteria enriched near the wellhead, and for a distally abundant OTU
99 belonging to the Marine Group I Archaea (OTU-15; Fig. 2). These data, along with
100 mRNA expression profiles (Dataset S1), imply spill-responsive bacterial communities
101 were still viable and active ≥ 75 days after well closure.

102 Likewise, bacterial genome bins were highly and exponentially enriched near MC252
103 (Fig. S4, Table S3); the average genome read-coverage was only 0.7 to 6.2 \times across
104 widespread sites 10-60 km away, but was 7.3 \times to 42.6 \times among sites within 2.7 km of
105 MC252 (Fig. 1, Fig. S5). The genomes were found across all near-well sites, as well as at
106 many less impacted distal locations, depicting a remarkably uniform community response
107 across vast areas of the seafloor. Importantly, these oil responsive bacteria were
108 universally distributed, with detection of 54 to 100% of contigs from each of the 51 bins
109 across all 13 sites (Table S4), with 6 Gammaproteobacteria and 2 Bacteroidetes bins
110 present at sites spanning the entire 60 km (Table S5, Fig. S6). These data suggest that, in
111 addition to a few highly abundant OTUs previously identified in the plume and along the
112 seafloor^{10,15,18}, there was also a widespread community-level response. Bacteria at the

113 furthestmost outreaches of our study area were either responding to low levels of
114 hydrocarbon contamination or were background community members.

115 Many of the DWH seafloor genomes were phylogenetically clustered (52%, co-
116 binning excluded). Clusters in each of our sampled proteobacterial classes and the
117 Bacteroidetes shared average amino acid identities of 60-86%, which broadly equates to
118 genus or family level relatedness²⁵. Of these, the Gammaproteobacteria exhibited 5
119 distinct inter-bin clusters (Table S6). Prevalent phylogenetic clustering suggests strong
120 habitat selection for traits shared among genetically similar organisms²⁶.
121 Gammaproteobacterial clusters included genomes related to sulfur-oxidizing *Candidatus*
122 *Halobeggiatoa*²⁷ (Fig. 2), and to hydrocarbonoclastic *Colwellia*, *Cycloclasticus* and
123 *Porticoccus* (Cellvibrionales) species²⁸⁻³⁰ (Table S7, Fig. S7). Of these, *Colwellia* and
124 *Cycloclasticus* were typical plume genera¹⁴. When compared with cultivated
125 representatives, EMIRGE-reconstructed 16S rRNA gene sequences related to
126 *Cycloclasticus* and *Porticoccus* belonged to distinct DWH seafloor clades (Fig. S8),
127 while extremely diverse sequences were associated with the *Colwellia-Thalassomonas-*
128 *Glaciecola* group (Fig. S8).

129 The phylogenetic characteristics of the proximal sediment communities suggested a
130 strong potential for sulfur and hydrocarbon metabolism; also supported by gene content
131 and gene expression profiles. Among the most highly expressed genes were those
132 exhibiting significant differential expression between proximal and distal sites for sulfur,
133 hydrocarbon and nitrogen metabolism (Fig. 3). An increase in genes involved in N
134 metabolism was previously identified in these oil-polluted sediments^{10,31}. Our data show
135 these genes – involved in the denitrification pathway and nitrite/hydroxylamine
136 oxidation/reduction – were significantly up-regulated at proximal sites. While
137 denitrification-related activity was associated with several Gammaproteobacteria, it was
138 mostly linked to anaerobic sulfur oxidation by two *Halobeggiatoa*-like Thiotrichaceae
139 (GSC1 and 3; Fig. 3 and S9). These Thiotrichaceae also expressed of hydroxylamine
140 reductase genes (EC 1.7.2.6), which they likely used as a supplemental method for
141 reducing nitrite³² to ammonia, although a hydroxylamine reductase required for the final
142 step (hydroxylamine reduction to ammonia) was not detected in any Thiotrichaceae
143 genome bin. Concomitant up-regulation of oxidative phosphorylation genes reflects the
144 classic oscillating aerobic-anaerobic lifestyle of this group of bacteria³³. Genes involved
145 in Thiotrichaceae sulfur oxidation, and deltaproteobacterial sulfate reduction, were both
146 significantly up-regulated near MC252, which implies an active spill-promoted sulfur
147 cycle formed in the top 1 cm of seafloor sediment.

148 Genes indicative of hydrocarbon degradation (HCD) were concentrated in the
149 Gammaproteobacteria (179 genes or 66%; Table 1 and Table S8). Half of the bacterial
150 genomes (n=25) contained genes associated with the degradation of hydrocarbons, most
151 belonging to aerobic pathways (Table S8). These genes were observed almost exclusively
152 in additive combinations targeting: (1) *n*-alkanes, (2) *n*-alkanes + aromatics, or (3) *n*-
153 alkanes + aromatics + PAHs (Figs S4 and S9), whereby alkane degradation potential is
154 the common denominator. As such, genomes with genes for *n*-alkane degradation were
155 cumulatively the most abundant near MC252 (Fig. S4). We observed the expression of
156 several genes involved in the aerobic degradation of these 3 hydrocarbon classes across
157 multiple seafloor sites, although a greater proportion of genes targeting *n*-alkane versus
158 (poly)aromatic substrates were expressed at distal sites (Fig. 4). Overall a greater number

159 of genes associated with hydrocarbon degradation were up-regulated proximally (Fig. 3),
160 likely due to the higher average concentration of hydrocarbons near MC252 (Fig. S4)¹⁰.

161 The genes that were differentially expressed for aerobic HCD and associated
162 oxidative phosphorylation belonged to 2 largely distinct groups of Gammaproteobacteria
163 (Fig. 3). The first group exhibited higher expression of genes at proximal sites involved
164 in the degradation of short-to-mid chain length *n*-alkanes, aromatics and PAHs (*Ca.*
165 *Cycloclasticus* GSC8 and GSC9, *Ca.* Thiotrichales GSC21, gammaproteobacterium
166 GSC16, and Chromatiales/Thiotrichales-related GSC22). The second up-regulated genes
167 at distal sites associated with short-to-mid chain *n*-alkanes (*Ca.* Colwellia GSC7, and *Ca.*
168 Cellvibrionales GSC14). While both groups of bacteria were active across wide
169 distances, the second group appeared to be more active at 10 and 15 km distances (Fig. 4)
170 where total petroleum hydrocarbon and *n*-alkane concentrations were generally lower¹⁰
171 (Fig. S4).

172 We identified 3 candidate PAH degraders (*Ca.* *Cycloclasticus* GSC9,
173 Chromatiales/Thiotrichales-related GSC22, and *Ca.* Cellvibrionales GSC15), of which
174 GSC9 and GSC22 exhibited active expression of PAH dioxygenases, particularly near-
175 well. Related *Cycloclasticus* species and *Porticoccus hydrocarbonoclasticus*
176 (Cellvibrionales) are known for their ability to degrade various PAHs^{28,29}. All 3 candidate
177 PAH degraders had broadly equivalent spatial abundances (Fig. S5), and each genome
178 has 17-24 subunits (large and small) of diverse ring-hydroxylating dioxygenases that,
179 except for 2, closely resemble dioxygenases used for the oxidation of PAHs and other
180 aromatic hydrocarbons (Fig. S9 and Dataset S1)³⁴. As previously observed in
181 *Cycloclasticus* genomes^{35,36}, each of our 3 DWH genomes had a greater proportion of
182 large over small PAH dioxygenase subunits (58-63% in genomes and 64% unbinned).

183 PAH dioxygenase sequences from the GSC9, GSC22 and GSC15 genomes chiefly
184 resembled naphthalene dioxygenases (23 large and 17 small subunits), while the
185 remainder (20 large and 11 small subunits) were more closely related to
186 biphenyl/benzene, anthranilate/ (ortho-halo)benzoate or pyrene dioxygenases (Fig. S10
187 and Table S8). PAH dioxygenases produce dihydrodiols³⁷. Canonical naphthalene,
188 anthranilate and pyrene *cis*-dihydrodiol dehydrogenases (EC 1.3.1.29, EC 1.3.1.49) were
189 not evident in our DWH genomes. However, all 3 genomes possess *cis*-2,3-
190 dihydrobiphenyl-2,3-diol dehydrogenase (*bphB*) and 2,3-dihydroxybiphenyl 1,2-
191 dioxygenase (*bphC*) genes. BphB (EC 1.3.1.56) is a multi-substrate enzyme that acts on
192 biphenyl-2,3-diol and a wide range of PAH dihydrodiols³⁷, including those relevant to
193 MC252 oil, such as naphthalene 1,2-dihydrodiol, and phenanthrene and chrysene 3,4-
194 dihydrodiols⁸. BphC degrades the catechol product of BphB. These bacteria appear to use
195 a universal pathway for the catabolism of early PAH degradation products, as opposed to
196 distinct dihydrodiol dehydrogenases per PAH substrate identified in a collection of Gulf
197 of Mexico seawater-associated Gammaproteobacteria³⁸.

198 Genetic mechanisms for BTEX degradation were previously found to be enriched at
199 highly contaminated seafloor sites¹⁰. Through assembly and bin assignment, we were
200 able to link genes used for toluene, xylene and benzene degradation with at least 4
201 Gammaproteobacteria: GSC9, GSC22, GSC16 (related to *n*-alkane-degrading
202 gammaproteobacterium HdN1³⁹) and GSC24 (related to Oceanospirillales *Hahella*
203 *chejuensis*). Collectively these 4 genomic bins represented 8% of the average genome
204 abundance at the proximal sites (Fig. S4). All 4 bins had genes encoding xylene

205 monooxygenase-like enzymes (Xyl), which oxidize toluene and xylenes⁴⁰. *xylAMM* genes
206 were expressed by GSC16 and was significantly higher at the proximal sites (Fig. 3). In
207 contrast, GSC9 demonstrated greater expression, albeit not significantly, of a gene (*tmoA*)
208 encoding part of a largely unbinned aerobic toluene-4-monooxygenase system at distal
209 sites, suggesting toluene/xylene metabolism very likely occurred across the seafloor,
210 although the organisms and mechanisms varied. GSC9 and GSC22 also had genes
211 encoding multicomponent phenol hydroxylase like enzymes (Dmp), which oxidize
212 phenol, benzene and toluene⁴¹. Further to these mechanisms, the 3 candidate PAH
213 degraders (GSC9, GSC22 and GSC15) had putative benzene 1,2-dioxygenases (also
214 similar to biphenyl dioxygenases) and catechol 2,3-dioxygenases (EC 1.13.11.2),
215 suggesting these taxa could generate catechol by benzene oxidation, which could then be
216 converted into 2-hydroxymuconate-semialdehyde, and sequentially transformed into
217 pyruvate (Fig. S9). While hydrocarbon degradation mechanisms identified in these
218 surface sediment communities were overwhelmingly aerobic, there were a few
219 exceptions in sequence data not binned to genomes. These include a single set of
220 anaerobic ethylbenzene dehydrogenase genes (*ebdACBA*); and benzylsuccinate synthase
221 genes (*bssCAB* and *bssCA*), which can be used for anaerobic toluene oxidation⁴², and
222 were observed in the lower anaerobic layer of seafloor sediments polluted with oil from
223 MC252⁴³.

224 Widespread evidence for alkane oxidation was associated with 3 different
225 mechanisms that target gaseous C2-C4 short-chain and liquid C5-C10 mid-chain alkanes.
226 Genes associated with the oxidation of both short to mid length *n*-alkane groups were
227 expressed across at least 34 km of the Gulf of Mexico seafloor (Fig. 4). Of these, Alk and
228 CYP153 enzymes act on mid-chain alkanes from pentane to decane (C5-C10)⁴⁴.
229 Transmembrane 1-alkane monooxygenase (AlkB ± AlkGT rubredoxin/rubredoxin
230 reductase) and membrane-bound cytochrome P450 CYP153 (± ferredoxin/ferredoxin
231 reductase) hydroxylases genes were both present, although CYP153 were more
232 commonly expressed. Pathway analysis of 14 key gammaproteobacterial bins with alkane
233 hydroxylases suggests 1-alcohol generated by alkane oxidation could be converted
234 sequentially to aldehydes by alcohol dehydrogenase (EC 1.1.1.1), carboxylates by
235 aldehyde dehydrogenase (NAD, EC 1.2.1.3), and acetyl-CoA via beta-oxidation (Fig. 3).

236 Also present were genes resembling particulate and soluble methane or ammonia
237 monooxygenases. These constitute a group of related multi-substrate enzymes that
238 preferentially target methane or ammonia⁴⁵. They can be used by methanotrophs, in the
239 absence of methane, to oxidize short *n*-alkanes, namely C2-C4 gases (ethane, butane,
240 propane) and C5 liquid (pentane)^{46,47}. Longer C6-C8 alkanes, can also be used, albeit at
241 an appreciably slower rate⁴⁶. In comparison, *Mycobacterium* strains preferentially use
242 soluble monooxygenases to oxidize alkanes⁴⁸. We recovered a diverse group of 5
243 genomes with particulate methane monooxygenase like genes (*Candidates* Colwellia
244 GSC7, Cycloclasticus GSC8, Cellvibrionales GSC14 and Thiotrichales GSC21, and
245 gammaproteobacterium IMCC2047 relative GSC18). Thiotrichales GSC21 also had
246 genes encoding the components of a soluble monooxygenase (sMMO), which is used in
247 place of the particulate enzyme (pMMO) under copper limiting conditions⁴⁹. All 5
248 genomes lack evidence for methanol or hydroxylamine oxidation, suggesting an inability
249 to utilize the products of methane or ammonia oxidation. However, all had the genetic
250 capacity to convert 1-alcohols generated from *n*-alkane oxidation to acetyl-CoA (Fig.

251 S9). We therefore predict that these bacteria primarily used pMMO \pm sMMO to oxidize
252 *n*-alkanes. The lack of dedicated methanotrophs plausibly reflects the absence of trapped
253 methane in the post-spill sediments, and is consistent with the absence of methane in the
254 late-stage plume⁵⁰.

255 We found that both types of methane monooxygenase like genes were expressed
256 across multiple seafloor locations; three bacteria expressed particulate *pmo* genes (GSC7,
257 GSC8, GSC14), while Thiotrichales GSC21 expressed soluble *mmo* genes (Fig. 4).
258 GSC21 co-expressed a short chain alcohol dehydrogenase gene, resembling a NADH-
259 dependent butanol dehydrogenase gene (*bdhA*), which it may have used to metabolize 1-
260 alcohol produced by alkane degradation (Fig. S9). We detected expression of parts of the
261 downstream alkane degradation pathway by GSC7 and GSC8 (Dataset S1). All genes
262 comprising the pathway from 1-alcohol to the first steps in the beta-oxidation pathway
263 were expressed by GSC14, including multiple NAD-dependent alcohol dehydrogenase
264 (EC 1.1.1.1) and aldehyde dehydrogenase genes.

265

266 Conclusions

267 Gulf of Mexico microorganisms are naturally exposed to oil seeps^{5,13} and frequent spills².
268 Our genomic and metabolic reconstruction of oil-impacted communities distributed
269 across the seafloor indicates that a large common collection of bacteria responded to the
270 DWH spill, many of which possessed hydrocarbonoclastic potential. A
271 large degree of apparent functional redundancy among HCD strategies suggests that the
272 Gulf of Mexico harbors functionally robust communities that are well poised to take
273 advantage of petroleum hydrocarbon influxes. We observed a strong environmental
274 selection preference for genetically similar organisms, implying that the preservation or
275 sharing of opportunistic hydrocarbonoclastic (and S-oxidizing) traits was important
276 among these DWH organisms. Due to the substrate promiscuity of many hydrocarbon-
277 degrading enzymes^{34,41,44,48}, it is unclear whether actively transcribed genes resulted in
278 competition for the same substrates or niche differentiation. Nevertheless, our results
279 show that several closely related hydrocarbon-degrading genes were concomitantly
280 expressed, and that individual bacterial populations appeared to occupy more than one
281 niche by co-utilizing functionally distinct hydrocarbon-degrading genes.

282

283 Methods

284 **Sampling and nucleic acid sequencing.** Thirteen seafloor sediment cores were collected
285 between 28 September and 19 October 2010 at radially distributed locations around the
286 capped MC252 wellhead (x7 cores between 0.3 and 2.7 km from MC252), and along a
287 distal southwesterly linear transect (x6 cores between 10.1 and 59.5 km from MC252)
288 (Fig. S1 and Table S1)¹⁰. The outer surfaces of 0 to 1 cm deep cores were removed prior
289 to DNA and RNA extractions¹⁰. DNA extraction and whole genome shotgun (WGS)
290 sequencing are described by Mason et al.¹⁰. Briefly, DNA extracted from the cores was
291 fragmented and prepared for sequencing using Illumina's TruSeq DNA Sample Prep Kit.
292 Each library was sequenced on a full HiSeq2000 lane at the Institute for Genomics and
293 Systems Biology's Next Generation Sequencing Core (Argonne National Laboratory).
294 This yielded ~18 Gb of sequence per sample with 2 x 101 bp reads and insert sizes of
295 ~135 bp. Low quality reads were trimmed using Sickle v. 1.29 with a quality score

296 threshold of Q=3, or removed if trimmed to <80 bp long
297 (<https://github.com/najoshi/sickle>).

298 RNA was extracted in duplicate from cores using 0.5 g of sediment for each replicate.
299 A modified hexadecyltrimethylammonium bromide (CTAB) extraction buffer was used
300 as previously described⁵¹. Duplicate extracts were pooled, and purification was carried
301 out using the Qiagen AllPrep DNA/RNA Mini Kit with on-column DNase digestion
302 using the RNase-Free DNase Set (Qiagen, Valencia, CA). RNA from samples with low
303 yields (<150 ng RNA: 0.3 km, 1.1 km, 10.1 km, 15.1 km and 33.9 km samples) was
304 amplified using the Ambion MessageAmp II aRNA Amplification Kit (Foster City, CA).
305 RNA was converted into double stranded cDNA using the SuperScript Double-
306 Stranded cDNA Synthesis Kit with random hexamers (Invitrogen, Carlsbad, CA).
307 Double stranded cDNA was prepared for sequencing using the TruSeq Nano DNA kit
308 (Illumina, San Diego, CA). Prepared libraries (~440 bp long) were sequenced using an
309 Illumina HiSeq 2500 at the Yale Center for Genomic Analysis, with 3 libraries per lane,
310 and generating 150 bp paired-end reads. Adapter sequences were removed using
311 Cutadapt⁵², and reads were trimmed with Trimmomatic⁵³ (sliding window quality score
312 ≥15) and removed if shorter than 60 bp.

313
314 **Metagenome assembly.** Metagenomic sequences from each sample were first assembled
315 individually using the IDBA-UD v. 1.1.0 metagenome assembler⁵⁴. Consolidation and
316 improved genome and HCD gene recovery was achieved via a co-assembly of 3
317 representative samples (0.5, 0.7, 0.9 km from MC252), due to high compositional
318 similarity among near-well communities. IDBA-UD was used for the co-assembly with
319 an optimal kmer range of 45 to 75 and step size of 15. Improved recovery of the highly
320 abundant GSC11 genome was achieved through selective re-assembly from the 0.5 km
321 metagenome using Velvet (kmer size = 63, expected kmer coverage = 147, kmer
322 coverage range = 92 to 225)⁵⁵.

323
324 **Genome binning, annotation, completion estimates and comparisons.** To bin contigs
325 ≥2 kbp long, we used the multi-parameter approach previously described by Handley et
326 al.⁵⁶. To better separate closely related genomes using emergent self-organizing maps
327 (ESOM), contig tetranucleotide frequencies were augmented with the coverage of that
328 contig in each spatially distinct sample (Fig. S11) based on the approach of Sharon et
329 al.²². The differential coverage of contigs was determined by mapping reads to co-
330 assembled contigs using bowtie2⁵⁷. Contig coverages per sample were scaled to 1 for
331 ESOM. Genome bin abundance heat and line plots were created using gplots and ggplot2
332 packages in R, respectively.

333 Contigs were annotated using the Integrated Microbial Genomes (IMG) pipeline⁵⁸,
334 and the Rapid Annotations using Subsystems Technology (RAST) Server⁵⁹. Predicted
335 protein sequences of potential HCD genes were searched against the National Center for
336 Biotechnology Information's (NCBI's) Conserved Domain Database⁶⁰.

337 Genome completion was estimated based on the presence of 107 single copy core
338 genes⁶¹, excluding *glyS*, *proS*, *pheT* and *rpoC*, which were missing or poorly recovered in
339 this study and according to Albertson et al.⁶². Core genes were detected using HMMER3
340 with the default cutoff⁶³. Estimates were similar to those obtained using AMPHORA2
341 with a set of 31 universal bacterial protein-coding housekeeping genes⁶⁴.

342 Genomes were compared using average amino acid identities (AAIs) including only
343 BLASTp matches that shared $\geq 30\%$ identity over an alignable region of $\geq 70\%$ sequence
344 length. 16S rRNA gene sequence and RecA and PAH dioxygenase predicted protein
345 sequences were compared using MEGA6⁶⁵ ClustalW alignments and neighbor-joining or
346 maximum-likelihood trees.

347
348 **Transcriptome read mapping.** Between 21 and 58 million trimmed paired-end reads
349 were mapped to the co-assembly using Bowtie2⁵⁷ with --end-to-end and --sensitive
350 settings. Alignments were sorted with Samtools⁶⁶. Hits were enumerated and filtered
351 using htseq-count in HTSeq v. 0.6.1p1⁶⁷ with default settings. Counts were normalized to
352 gene length and reads per sample by a modification of the approach described by
353 Mortazavi et al.⁶⁸, whereby normalization was to Reads Per Kilobase per Average library
354 size (RPKA; average = 44 million). To determine whether gene expression was up or
355 down regulated spatially, RPKA values were also normalized to the genome coverage in
356 each sample (RPKAC, as we required that at least one sample per gene had at least 10
357 mapped reads (un-normalized). To compare differential gene expression between
358 proximal and distal sites we used edgeR⁶⁹ on un-normalized data with RPKAC values
359 supplied as an offset matrix of correction factors to the generalized linear model. Sample
360 sizes were supplied as total library sizes.

361
362 **Ribosomal RNA sequence assembly and clustering.** To investigate beta diversity near
363 full-length small subunit (SSU) rRNA (gene) sequences were reconstructed using the
364 reference-guided Expectation Maximization Iterative Reconstruction of Genes from the
365 Environment (EMIRGE) method²³. WGS samples were rarefied to an equal depth of 64
366 million paired-end reads. Transcriptome samples were rarefied to between 22 and 32
367 million reads after first pre-selecting SSU rRNA specific transcriptome kmers using
368 bbduk (<http://sourceforge.net/projects/bbmap>) and the SILVA SSU rRNA database⁷⁰.
369 Sequences were reconstructed over 80 iterations using EMIRGE with the SILVA 111
370 SSU rRNA database. Gapped and chimeric sequences were removed using 64-bit
371 USEARCH v. 8.0 with the RDP Gold v. 9 database (<http://drive5.com/>). RNA and RNA
372 gene sequences were co-clustered at $\geq 97\%$ similarity into operational taxonomic units
373 (OTUs) using -cluster_otus (32-bit USEARCH v. 8.1)⁷¹ after sorting by size. OTU
374 representative sequences were identified by 64-bit USEARCH global alignment to the
375 full SILVA 111 SSU rRNA database, and by RDP classifier v. 2.6⁷² with a 0.8 bootstrap
376 cutoff.

377
378 **Sequence accession.** Metagenomic and EMIRGE assemblies, and transcriptome reads
379 are accessible via NCBI BioProject's PRJNA258478 and PRJNA342256, respectively.
380 Metagenome annotations are accessible via IMG ID 3300003691.

381
382 **Acknowledgements**

383 This work was supported by Alfred P. Sloan Foundation and Exxon-Mobile grants
384 awarded to JAG, and a Royal Society of NZ Rutherford Discovery Fellowship awarded
385 to KMH. Partial support was provided by the U.S. Dept. of Energy under contracts DE-
386 AC02-06CH11357 (ANL) and DE-AC05-76RL01830 (PNNL). We thank Christian
387 Sieber (JGI) for transcriptome rRNA sequence assembly, and acknowledge resources

388 provided by the University of Chicago Research Computing Center, NERSC, and the
389 University of Auckland NeSI high-performance computing facilities and Centre for
390 eResearch.

391

392 **References**

- 393 1 Barron, M. G. Ecological impacts of the deepwater horizon oil spill: implications
394 for immunotoxicity. *Toxicol Pathol* **40**, 315-320, doi:10.1177/0192623311428474
395 (2012).
- 396 2 USCG. Polluting Incidents In and Around U.S. Waters. A Spill/Release
397 Compendium: 1969 - 2011. (2012).
- 398 3 Peterson, C. H. *et al.* A tale of two spills: Novel science and policy implications
399 of an emerging new oil spill model. *BioScience* **461**, 461-469 (2012).
- 400 4 Atlas, R. M. & Hazen, T. C. Oil biodegradation and bioremediation: a tale of the
401 two worst spills in U.S. history. *Environ Sci Technol* **45**, 6709-6715,
402 doi:10.1021/es2013227 (2011).
- 403 5 MacDonald, I. *et al.* Natural and unnatural oil slicks in the Gulf of Mexico. *J.*
404 *Geophys. Res. Oceans* **120**, 8364–8380, doi:doi:10.1002/2015JC011062 (2015).
- 405 6 Camilli, R. *et al.* Tracking hydrocarbon plume transport and biodegradation at
406 Deepwater Horizon. *Science* **330**, 201-204, doi:10.1126/science.1195223 (2010).
- 407 7 Valentine, D. L. *et al.* Fallout plume of submerged oil from Deepwater Horizon.
408 *Proc Natl Acad Sci U S A* **111**, 15906-15911, doi:10.1073/pnas.1414873111
409 (2014).
- 410 8 Zukunft, P. Operational Science Advisory Team Report (OSAT). (2010).
- 411 9 Yan, B. *et al.* Sustained deposition of contaminants from the Deepwater Horizon
412 spill. *Proc Natl Acad Sci U S A* **113**, E3332-3340, doi:10.1073/pnas.1513156113
413 (2016).
- 414 10 Mason, O. U. *et al.* Metagenomics reveals sediment microbial community
415 response to Deepwater Horizon oil spill. *ISME J* **8**, 1464-1475,
416 doi:10.1038/ismej.2013.254 (2014).
- 417 11 P.H. Pritchard, J.G. Mueller, J.C. Rogers, F.V. Kremer & Glaser, J. A. Oil spill
418 bioremediation: experiences, lessons and results from the Exxon Valdez oil spill
419 in Alaska. *Biodegradation* **3**, 315-335 (1992).
- 420 12 Brooijmans, R. J., Pastink, M. I. & Siezen, R. J. Hydrocarbon-degrading bacteria:
421 the oil-spill clean-up crew. *Microb Biotechnol* **2**, 587-594, doi:10.1111/j.1751-
422 7915.2009.00151.x (2009).
- 423 13 Joye, S. B., Teske, A. P. & Kostka, J. E. Microbial Dynamics Following the
424 Macondo Oil Well Blowout across Gulf of Mexico Environments. *BioScience* **64**,
425 766-777, doi:10.1093/biosci/biu121 (2014).
- 426 14 Redmond, M. C. & Valentine, D. L. Natural gas and temperature structured a
427 microbial community response to the Deepwater Horizon oil spill. *Proc Natl Acad*
428 *Sci USA* **109**, 20292-20297, doi:10.1073/pnas.1108756108 (2012).
- 429 15 Hazen, T. C. *et al.* Deep-sea oil plume enriches indigenous oil-degrading bacteria.
430 *Science* **330**, 204-208, doi:10.1126/science.1195979 (2010).
- 431 16 King, G. M., Smith, C. B., Tolar, B. & Hollibaugh, J. T. Analysis of composition
432 and structure of coastal to mesopelagic bacterioplankton communities in the

- 433 northern gulf of Mexico. *Front Microbiol* **3**, 438, doi:10.3389/fmicb.2012.00438
434 (2013).
- 435 17 Rivers, A. R. *et al.* Transcriptional response of bathypelagic marine
436 bacterioplankton to the Deepwater Horizon oil spill. *ISME J* **7**, 2315-2329,
437 doi:10.1038/ismej.2013.129 (2013).
- 438 18 Mason, O. U. *et al.* Metagenome, metatranscriptome and single-cell sequencing
439 reveal microbial response to Deepwater Horizon oil spill. *ISME J* **6**, 1715-1727,
440 doi:10.1038/ismej.2012.59 (2012).
- 441 19 Mason, O. U., Han, J., Woyke, T. & Jansson, J. K. Single-cell genomics reveals
442 features of a *Colwellia* species that was dominant during the Deepwater Horizon
443 oil spill. *Front Microbiol* **5**, 332, doi:10.3389/fmicb.2014.00332 (2014).
- 444 20 Spier, C., Stringfellow, W. T., Hazen, T. C. & Conrad, M. Distribution of
445 hydrocarbons released during the 2010 MC252 oil spill in deep offshore waters.
446 *Environ Pollut* **173**, 224-230, doi:10.1016/j.envpol.2012.10.019 (2013).
- 447 21 Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence
448 signatures. *Genome Biol* **10**, R85, doi:10.1186/gb-2009-10-8-r85 (2009).
- 449 22 Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in
450 bacterial species, strains, and phage during infant gut colonization. *Genome Res*
451 **23**, 111-120, doi:10.1101/gr.142315.112 (2013).
- 452 23 Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W. & Banfield, J. F.
453 EMIRGE: reconstruction of full-length ribosomal genes from microbial
454 community short read sequencing data. *Genome Biol* **12**, R44, doi:10.1186/gb-
455 2011-12-5-r44 (2011).
- 456 24 Blazewicz, S. J., Barnard, R. L., Daly, R. A. & Firestone, M. K. Evaluating rRNA
457 as an indicator of microbial activity in environmental communities: limitations
458 and uses. *ISME J* **7**, 2061-2068, doi:10.1038/ismej.2013.102 (2013).
- 459 25 Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for
460 prokaryotes. *J Bacteriol* **187**, 6258-6264, doi:10.1128/JB.187.18.6258-6264.2005
461 (2005).
- 462 26 Horner-Devine, M. C. & Bohannon, B. J. Phylogenetic clustering and
463 overdispersion in bacterial communities. *Ecology* **87**, S100-108 (2006).
- 464 27 Grünke, S. *et al.* Mats of psychrophilic thiotrophic bacteria associated with cold
465 seeps of the Barents Sea. *Biogeosciences* **9**, 2947-2960, doi:10.5194/bg-9-2947-
466 2012 (2012).
- 467 28 Geiselbrecht, A. D., Hedlund, B. P., Tichi, M. A. & Staley, J. T. Isolation of
468 Marine Polycyclic Aromatic Hydrocarbon (PAH)-Degrading Cycloclasticus
469 Strains from the Gulf of Mexico and Comparison of Their PAH Degradation
470 Ability with That of Puget Sound Cycloclasticus Strains. *Appl Environ Microbiol*
471 **64**, 4703-4710 (1998).
- 472 29 Gutierrez, T., Nichols, P. D., Whitman, W. B. & Aitken, M. D. *Porticoccus*
473 *hydrocarbonoclasticus* sp. nov., an aromatic hydrocarbon-degrading bacterium
474 identified in laboratory cultures of marine phytoplankton. *Appl Environ Microbiol*
475 **78**, 628-637, doi:10.1128/AEM.06398-11 (2012).
- 476 30 Baelum, J. *et al.* Deep-sea bacteria enriched by oil and dispersant from the
477 Deepwater Horizon spill. *Environ Microbiol* **14**, 2405-2416, doi:10.1111/j.1462-
478 2920.2012.02780.x (2012).

- 479 31 Scott, N. M. *et al.* The microbial nitrogen cycling potential is impacted by
480 polyaromatic hydrocarbon pollution of marine sediments. *Front Microbiol* **5**, 108,
481 doi:10.3389/fmicb.2014.00108 (2014).
- 482 32 MacGregor, B. J. *et al.* Why orange Guaymas Basin Beggiatoa spp. are orange:
483 single-filament-genome-enabled identification of an abundant octaheme
484 cytochrome with hydroxylamine oxidase, hydrazine oxidase, and nitrite reductase
485 activities. *Appl Environ Microbiol* **79**, 1183-1190, doi:10.1128/AEM.02538-12
486 (2013).
- 487 33 Schulz, H. N. & Jorgensen, B. B. Big bacteria. *Annu Rev Microbiol* **55**, 105-137,
488 doi:10.1146/annurev.micro.55.1.105 (2001).
- 489 34 Jouanneau, Y., Meyer, C., Jakoncic, J., Stojanoff, V. & Gaillard, J.
490 Characterization of a naphthalene dioxygenase endowed with an exceptionally
491 broad substrate specificity toward polycyclic aromatic hydrocarbons.
492 *Biochemistry* **45**, 12380-12391, doi:10.1021/bi0611311 (2006).
- 493 35 Lai, Q., Li, W., Wang, B., Yu, Z. & Shao, Z. Complete genome sequence of the
494 pyrene-degrading bacterium *Cycloclasticus* sp. strain P1. *J Bacteriol* **194**, 6677,
495 doi:10.1128/JB.01837-12 (2012).
- 496 36 Cui, Z., Xu, G., Li, Q., Gao, W. & Zheng, L. Genome Sequence of the Pyrene-
497 and Fluoranthene-Degrading Bacterium *Cycloclasticus* sp. Strain PY97M.
498 *Genome Announc* **1**, doi:10.1128/genomeA.00536-13 (2013).
- 499 37 Jouanneau, Y. & Meyer, C. Purification and characterization of an arene cis-
500 dihydrodiol dehydrogenase endowed with broad substrate specificity toward
501 polycyclic aromatic hydrocarbon dihydrodiols. *Appl Environ Microbiol* **72**, 4726-
502 4734, doi:10.1128/AEM.00395-06 (2006).
- 503 38 Dombrowski, N. *et al.* Reconstructing metabolic pathways of hydrocarbon-
504 degrading bacteria from the Deepwater Horizon oil spill. *Nat Microbiol*, 16057,
505 doi:10.1038/nmicrobiol.2016.57 (2016).
- 506 39 Zedelius, J. *et al.* Alkane degradation under anoxic conditions by a nitrate-
507 reducing bacterium with possible involvement of the electron acceptor in
508 substrate activation. *Environ Microbiol Rep* **3**, 125-135, doi:10.1111/j.1758-
509 2229.2010.00198.x (2011).
- 510 40 Harayama, S. *et al.* Characterization of five genes in the upper-pathway operon of
511 TOL plasmid pWW0 from *Pseudomonas putida* and identification of the gene
512 products. *J Bacteriol* **171**, 5048-5055 (1989).
- 513 41 Cafaro, V., Notomista, E., Capasso, P. & Di Donato, A. Regiospecificity of two
514 multicomponent monooxygenases from *Pseudomonas stutzeri* OX1: molecular
515 basis for catabolic adaptation of this microorganism to methylated aromatic
516 compounds. *Appl Environ Microbiol* **71**, 4736-4743,
517 doi:10.1128/AEM.71.8.4736-4743.2005 (2005).
- 518 42 Beller, H. R. & Spormann, A. M. Analysis of the novel benzylsuccinate synthase
519 reaction for anaerobic toluene activation based on structural studies of the
520 product. *J Bacteriol* **180**, 5454-5457 (1998).
- 521 43 Kimes, N. E. *et al.* Metagenomic analysis and metabolite profiling of deep-sea
522 sediments from the Gulf of Mexico following the Deepwater Horizon oil spill.
523 *Front Microbiol* **4**, 50, doi:10.3389/fmicb.2013.00050 (2013).

- 524 44 van Beilen, J. B. *et al.* Cytochrome P450 alkane hydroxylases of the CYP153
525 family are common in alkane-degrading eubacteria lacking integral membrane
526 alkane hydroxylases. *Appl Environ Microbiol* **72**, 59-65,
527 doi:10.1128/AEM.72.1.59-65.2006 (2006).
- 528 45 O'Neill, J. G. & Wilkinson, J. F. Oxidation of ammonia by methane-oxidizing
529 bacteria and the effects of ammonia on methane oxidation. *Journal of General*
530 *Microbiology* **100**, 407-412 (1977).
- 531 46 Colby, J., Stirling, D. I. & Dalton, H. The Soluble Methane Mono-oxygenase of
532 *Methylococcus capsulatus* (Bath). *Biochem J* **165**, 395-402 (1977).
- 533 47 Patel, R. N., Hou, C. T., Laskin, A. I., Felix, A. & Derelanko, P. Microbial
534 oxidation of gaseous hydrocarbons. II. Hydroxylation of alkanes and epoxidation
535 of alkenes by cell-free particulate fractions of methane-utilizing bacteria. *J*
536 *Bacteriol* **139**, 675-679 (1979).
- 537 48 Martin, K. E., Ozsvar, J. & Coleman, N. V. SmoXYB1C1Z of *Mycobacterium* sp.
538 strain NBB4: a soluble methane monooxygenase (sMMO)-like enzyme, active on
539 C2 to C4 alkanes and alkenes. *Appl Environ Microbiol* **80**, 5801-5806,
540 doi:10.1128/AEM.01338-14 (2014).
- 541 49 Nielsen, A. K., Gerdes, K. & Murrell, J. C. Copper-dependent reciprocal
542 transcriptional regulation of methane monooxygenase genes in *Methylococcus*
543 *capsulatus* and *Methylosinus trichosporium*. *Mol Microbiol* **25**, 399-409 (1997).
- 544 50 Kessler, J. D. *et al.* A persistent oxygen anomaly reveals the fate of spilled
545 methane in the deep Gulf of Mexico. *Science* **331**, 312-315,
546 doi:10.1126/science.1199697 (2011).
- 547 51 DeAngelis, K. M., Silver, W. L., Thompson, A. W. & Firestone, M. K. Microbial
548 communities acclimate to recurring changes in soil redox potential status. *Environ*
549 *Microbiol* **12**, 3137-3149, doi:10.1111/j.1462-2920.2010.02286.x (2010).
- 550 52 Martin, M. Cutadapt removes adapter sequences from high-throughput
551 sequencing reads. *EMBnet.journal* **71**, 10-12,
552 doi:http://dx.doi.org/10.14806/ej.17.1.200 (2011).
- 553 53 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for
554 Illumina sequence data. *Bioinformatics* **30**, 2114-2120,
555 doi:10.1093/bioinformatics/btu170 (2014).
- 556 54 Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a de novo assembler
557 for single-cell and metagenomic sequencing data with highly uneven depth.
558 *Bioinformatics* **28**, 1420-1428, doi:10.1093/bioinformatics/bts174 (2012).
- 559 55 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly
560 using de Bruijn graphs. *Genome Res* **18**, 821-829, doi:10.1101/gr.074492.107
561 (2008).
- 562 56 Handley, K. M. *et al.* Biostimulation induces syntrophic interactions that impact
563 C, S and N cycling in a sediment microbial community. *ISME J* **7**, 800-816,
564 doi:10.1038/ismej.2012.148 (2013).
- 565 57 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat*
566 *Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
- 567 58 Markowitz, V. M. *et al.* IMG/M: the integrated metagenome data management
568 and comparative analysis system. *Nucleic Acids Res* **40**, D123-129,
569 doi:10.1093/nar/gkr975 (2012).

- 570 59 Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems
571 technology. *BMC Genomics* **9**, 75, doi:10.1186/1471-2164-9-75 (2008).
- 572 60 Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic
573 Acids Res* **43**, D222-226, doi:10.1093/nar/gku1221 (2015).
- 574 61 Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated
575 marine bacterial lineage. *ISME J* **6**, 1186-1199, doi:10.1038/ismej.2011.189
576 (2012).
- 577 62 Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by
578 differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533-
579 538, doi:10.1038/nbt.2579 (2013).
- 580 63 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195,
581 doi:10.1371/journal.pcbi.1002195 (2011).
- 582 64 Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences
583 with AMPHORA2. *Bioinformatics* **28**, 1033-1034,
584 doi:10.1093/bioinformatics/bts079 (2012).
- 585 65 Tamura, K., Stecher, G., Peterson, D., Filipinski, A. & Kumar, S. MEGA6:
586 Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725-
587 2729, doi:10.1093/molbev/mst197 (2013).
- 588 66 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
589 **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 590 67 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with
591 high-throughput sequencing data. *Bioinformatics* **31**, 166-169,
592 doi:10.1093/bioinformatics/btu638 (2015).
- 593 68 Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping
594 and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-
595 628, doi:10.1038/nmeth.1226 (2008).
- 596 69 Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor
597 package for differential expression analysis of digital gene expression data.
598 *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 (2010).
- 599 70 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data
600 processing and web-based tools. *Nucleic Acids Res* **41**, D590-596,
601 doi:10.1093/nar/gks1219 (2013).
- 602 71 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST.
603 *Bioinformatics* **26**, 2460-2461, doi:10.1093/bioinformatics/btq461 (2010).
- 604 72 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for
605 rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl
606 Environ Microbiol* **73**, 5261-5267, doi:10.1128/AEM.00062-07 (2007).
607

608

609 **Figure and table legends**

610

611 Figure 1. Average genome bin coverage and count (for bins with $>1 \times$ coverage) in the
612 co-assembly, and for the same collection of genome bins at each individual site; both
613 decrease noticeably with increasing distance from the wellhead, although some of the
614 genomes were still detectable 60 km away. Distance along the x-axis is not shown to
615 scale.

616

617 Figure 2. (A) The correlation of EMIRGE 16S rRNA gene and rRNA OTU abundances.
618 (B) The difference between proximal and distal sites. Read clockwise: higher rRNA gene
619 and rRNA in proximal locations (++); higher rRNA gene but lower rRNA proximally (+-
620); lower rRNA gene and rRNA proximally (--); and lower rRNA gene but higher rRNA
621 proximally (-+). (A-B) OTU numbers are given besides taxa points. Insets show the
622 highly abundant OTU 1. Based on abundance and phylogenetic affiliation, OTU 1
623 corresponds to the *Ca. Cellvibrionales* GSC11-15 genomes (93% identity to *P.*
624 *hydrocarbonoclasticus*). It also shares 100% identity (ID) with an iTag sequence ($>97\%$
625 similar to Greengenes OTU 248394) from a highly abundant uncultured
626 gammaproteobacterium previously identified in the contaminated near-well sediments¹⁰.
627 Thiotrichaceae OTU 5 corresponds to *Ca. Thiotrichaceae* GSC1 (99% ID to *Ca.*
628 *Halobeggiatoa* sp. HMW-S2528); *Cycloclasticus* OTUs 18 and 28 respectively
629 correspond to *Ca. Cycloclasticus* GSC8 and GSC9-10 (respectively 98% and 95% ID to
630 *Cycloclasticus zancles* 78-ME); *Colwellia* OTU 4 corresponds to *Ca. Colwellia* GSC4
631 and GSC9 (99% ID to *Colwellia psychrerythraea* 34H), and *Colwellia* OTU 16
632 corresponds to *Ca. Colwellia* GSC5-6 (97% ID to *Colwellia* sp. MT41).

633

634 Figure 3. Differentially expressed genes at proximal and distal sites associated with
635 hydrocarbon degradation, oxidative phosphorylation, and S cycling and N reduction
636 pathways. Genome bin numbers (GSC) are given beside each bar. Abbreviations:
637 degradation pathway (DP); beta oxidation (BO).

638

639 Figure 4. Spatial expression profiles of genes associated with hydrocarbon degradation.
640 Genome bin and gene identities are given for each row. *Significantly differentially
641 expressed genes at proximal (red font) versus distal (blue font) sites.

642

643 Table 1. Hydrocarbon degradation gene distributions.

644

645

646 **Supplementary figure, table and dataset legends**

647

648 Figure S1. (a) Location of well MC252, and (b) sampling locations. Samples >3 km from
649 MC252 are shown in blue, while near-well samples are shown in red (inset).

650

651 Figure S2. Coverage and GC content of contigs within each bin. Y-axes have different
652 scales (LHS of each plot). *Compositionally similar co-binned genomes with imprecise
653 coverage separation.

654

655 Figure S3. Phylogeny based on rRNA genes reconstructed from rarefied data, (a)
656 Phylogeny per site. Eukaryota are relatively abundant in distal sites (23-52% versus 6-
657 34% near-well). Bacteria, particularly Gammaproteobacteria increase in abundance
658 relative to both Eukaryota and Archaea. (b) Stacked bar chart of Gammaproteobacteria
659 rRNA gene sequence coverage per site based on RDP genus level designation or higher.
660 The average number of unique designations is 13 ± 2 (1 standard deviation) for near-well
661 sites, twice that for distal sites (7 ± 3). (c) Neighbor-joining tree depicting phylogenetic
662 16S rRNA gene diversity among the Gammaproteobacteria for near-well (red) and distal
663 samples (blue), compared with reference sequences (black). Genus (italics) and order
664 (bold) names are given for reference sequence clusters. Gammaproteobacterial richness at
665 near-well sites was 317 (15,000x total 16S rRNA gene coverage), and the richness at
666 distal sites was 152 (6,000x total 16S rRNA gene coverage).

667

668 Figure S4. (a) Boxplots of petroleum hydrocarbon and carbon concentrations in samples
669 0-3 km (near-well) and 10-60 km (distal) from the MC252 well-head. Abbreviations:
670 (total) petroleum hydrocarbons, (T)PHC; weight, wt. *Concentrations enriched near-well.
671 (b) Summed per site abundance of genome bins possessing genes associated with
672 hydrocarbon degradation. Points are fitted with exponential curves. Bin numbers are
673 listed in boxes central to each plot. Points in yellow represent the combined abundance of
674 bin numbers given in black and yellow.

675

676 Figure S5. Average genome bin coverages per site determined by mapping reads to the
677 co-assembly. Error bars = 1 standard deviation. Samples contributing to the co-assembly
678 (white points) are bolded and points shown in yellow. Plots are organized by
679 phylogenetic group: Gammaproteobacteria (green), Alphaproteobacteria (yellow),
680 Deltaproteobacteria (blue), Cytophaga-Flavobacterium-Bacteroides group (CFB, pink),
681 and unknown Bacteria (grey).

682

683 Figure S6. Heatmap and hierarchical clustering of genome bin abundance per site or Co-
684 assembly (Co). Samples contributing to the co-assembly are bolded. Abundances ≥ 10 are
685 assigned the same color (dark red). Bins (RHS) are colored by taxonomic group:
686 Gammaproteobacteria (green), Alphaproteobacteria (yellow), Deltaproteobacteria (blue),
687 Cytophaga-Flavobacterium-Bacteroides group (CFB, pink), and unknown Bacteria
688 (grey).

689

690 Figure S7. Maximum-Likelihood tree showing the phylogenetic relatedness among
691 recombinase A (RecA) predicted protein sequences from our seafloor genome bins

692 (green) and reference organisms (black). Reference sequence GenBank accession
693 numbers are in parentheses. Tree construction employed 190 amino acid positions and
694 500 bootstrap replicates.

695

696 Figure S8. Maximum-Likelihood trees showing the phylogenetic relatedness of rRNA
697 genes sequences from near-well (red) and distal (blue) sites that are similar to (a)
698 *Cycloclasticus*, (b) *Porticoccus*, and (c) *Colwellia* species. Sequence identifiers are in
699 parentheses. Norm Priors (NP) denote abundances per sample. Tree construction
700 employed near full-length 16S rRNA genes and 500 bootstrap replicates. 16S sequences
701 related to these hydrocarbonoclastic genera were recovered from rarified sequence data
702 up to ~10 km from MC252 for *Cycloclasticus* and *Porticoccus*, or 33.9 km in the case of
703 *Colwellia*.

704

705 Figure S9. (a) Idealized cell-schematic showing genes GSC1 likely uses to encode for
706 denitrification and sulfide oxidation. Genes present are in grey, while those absent are in
707 red. (b) Cell schematic of 14 gammaproteobacterial genome bins with the metabolic
708 potential to degrade hydrocarbons. Genes present are in black; those missing are in red.
709 Subunits are listed in order present in genomes, and separated by a dash if on different
710 contigs. Substrates (and their products) are categorized by color: C2-C10 alkanes (dark
711 blue), methane (light blue), aromatic hydrocarbons (orange), PAHs (purple), nitrogen
712 species (green), and other (black). All genome bins contain (near) complete beta
713 oxidation pathways that can be used to oxidize hexadecanoate or fatty acyl-CoA esters
714 intermediates to acetyl-CoA.

715

716 Figure S10. Maximum-Likelihood tree depicting the phylogenetic relatedness among
717 predicted protein sequences of (near) full length candidate PAH dioxygenases alpha
718 subunits from our genome bins (green) compared with reference naphthalene, pyrene and
719 anthranilate PAH dioxygenases and biphenyl, benzene and (ortho-halo-)benzoate
720 aromatic dioxygenases. GenBank accession numbers for reference sequences are in
721 parentheses. Tree construction employed 340 amino acid positions and 500 bootstrap
722 replicates.

723

724 Figure S11. (a) ESOM of co-assembly constructed using tetranucleotide frequencies of 5
725 kbp long genomic fragments and differential coverage. Bin numbers are shown in red.
726 Dark lines demark cluster edges. Collections of small clusters are contigs containing
727 virus-associated genes. (b-e) ESOM with data points representing 2 kbp long genomic
728 fragments and colored by bin (see key).

729

730 Table S1. Sequence data input (trimmed), and IDBA-UD assembly summary.

731

732 Table S2. Genomic bin characteristics.

733

734 Table S3. Pearson's correlations between distance and (hydro)carbon concentrations or
735 genome coverage.

736

737 Table S4. Percentage with mapped reads at each site.

738

739 Table S5. Average mapped-read genome bin coverages per site.

740

741 Table S6. Pairwise average amino acid identities (AAI) shared between genome bins.

742

743 Table S7. Average amino acid identities (AAI) shared between genome bins (left) and
744 reference genomes (top).

745

746 Table S8. Summary of candidate hydrocarbon degradation genes distributed among 25
747 bacterial genome bins, GSC52 and the unbinned fraction.

748

749 Dataset S1. Contigs per bin; key genes involved in hydrocarbon degradation, nitrate
750 reduction and sulfur oxidation; and mRNA read counts for key genes.

751

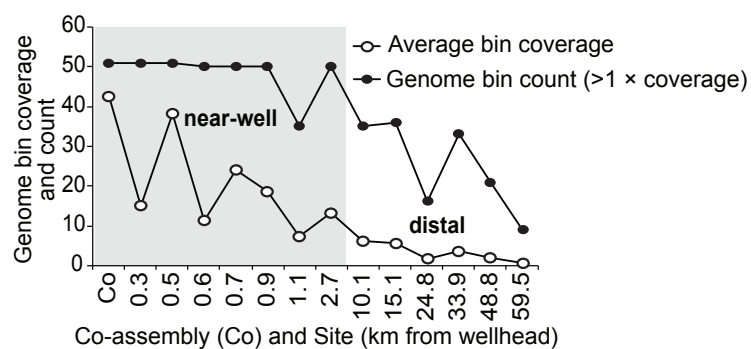


Figure 1. Average genome bin coverage and count (for bins with $>1 \times$ coverage) in the co-assembly, and for the same collection of genome bins at each individual site; both decrease noticeably with increasing distance from the wellhead, although some of the genomes were still detectable 60 km away. Distance along the x-axis is not shown to scale.

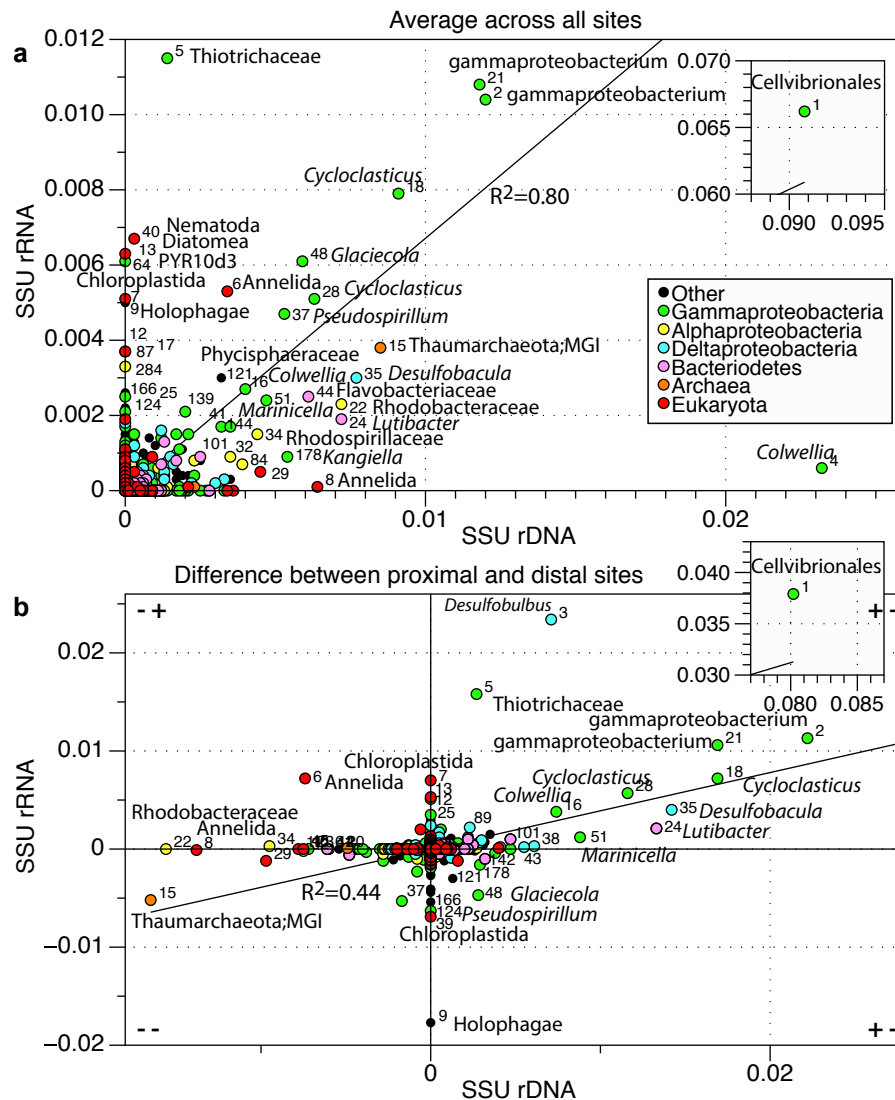


Figure 2. (A) The correlation of EMIRGE 16S rRNA gene and rRNA OTU abundances. (B) The difference between proximal and distal sites. Read clockwise: higher rRNA gene and rRNA in proximal locations (++); higher rRNA gene but lower rRNA proximally (+-); lower rRNA gene and rRNA proximally (--); and lower rRNA gene but higher rRNA proximally (-+). (A-B) OTU numbers are given besides taxa points. Insets show the highly abundant OTU 1. Based on abundance and phylogenetic affiliation, OTU 1 corresponds to the *Ca. Cellvibrionales* GSC11-15 genomes (93% identity to *P. hydrocarbonoclasticus*). It also shares 100% identity (ID) with an iTag sequence (>97% similar to Greengenes OTU 248394) from a highly abundant uncultured gammaproteobacterium previously identified in the contaminated near-well sediments¹⁰. Thiotrichaceae OTU 5 corresponds to *Ca. Thiotrichaceae* GSC1 (99% ID to *Ca. Halobeggiatoa* sp. HMW-S2528); *Cycloclasticus* OTUs 18 and 28 respectively correspond to *Ca. Cycloclasticus* GSC8 and GSC9-10 (respectively 98% and 95% ID to *Cycloclasticus zancles* 78-ME); *Colwellia* OTU 4 corresponds to *Ca. Colwellia* GSC4 and GSC9 (99% ID to *Colwellia psychrerythraea* 34H), and *Colwellia* OTU 16 corresponds to *Ca. Colwellia* GSC5-6 (97% ID to *Colwellia* sp. MT41).

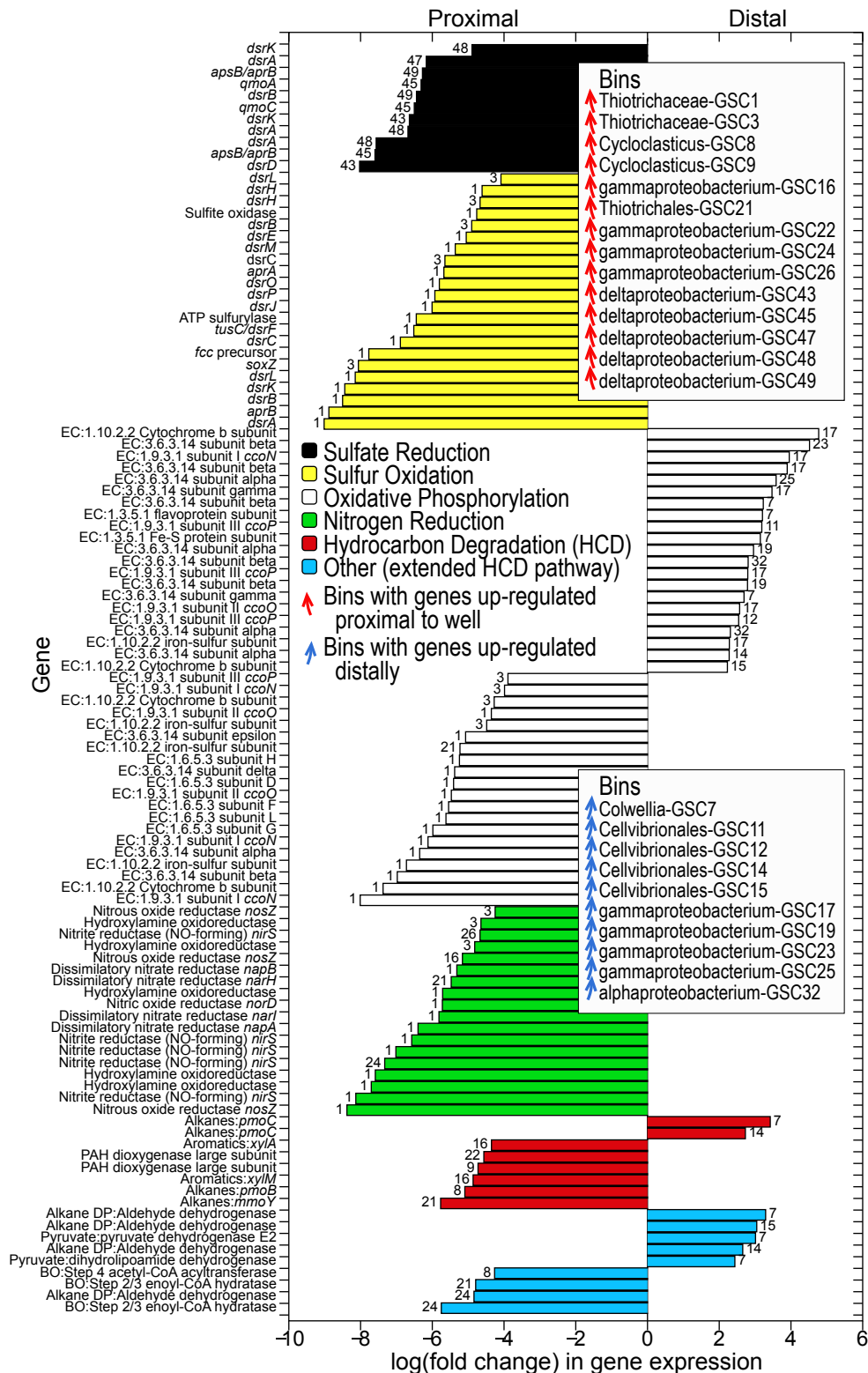


Figure 3. Differentially expressed genes at proximal and distal sites associated with hydrocarbon degradation, oxidative phosphorylation, and S cycling and N reduction pathways. Genome bin numbers (GSC) are given beside each bar. Abbreviations: degradation pathway (DP); beta oxidation (BO).

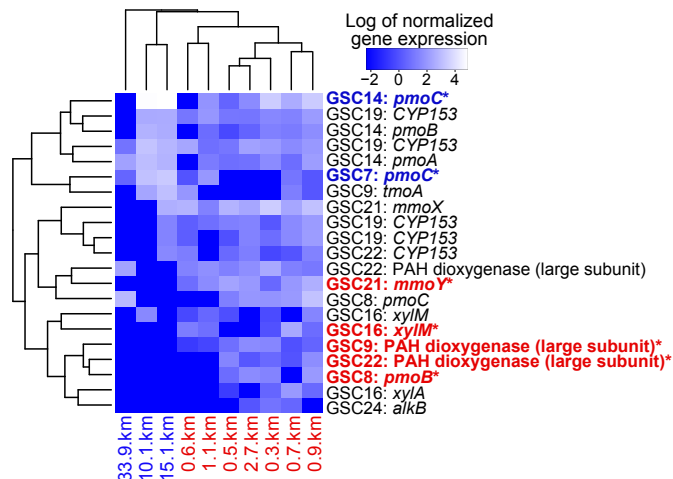


Figure 4. Spatial expression profiles of genes associated with hydrocarbon degradation. Genome bin and gene identities are given for each row. *Significantly differentially expressed genes at proximal (red font) versus distal (blue font) sites.

Table 1. Hydrocarbon degradation (HCD) gene distributions.

Group	HCD gene count	Relative abundance (%)	Normalized to CDS (%)*
Gamma proteobacteria	180	66	55
Unbinned	67	25	22
Delta proteobacteria	13	5	11
Bacteroidetes	9	3	7
Alpha proteobacteria	2	1	2
Phage associated	2	1	3
Total	273	100	100

*Normalized to total coding DNA sequence (CDS) per group.