Supporting Information

Integrated Omics Elucidate the Mechanisms Driving the Rapid Biodegradation of Deepwater Horizon Oil in Intertidal Sediments Undergoing Oxic-Anoxic Cycles

Smruthi Karthikeyan[1], Minjae Kim[1], Patrick Heritier-Robbins[1], Janet K. Hatt[1], Jim C. Spain[2], Will A. Overholt, Markus Huettel[3], Joel E. Kostka[4,5] and Konstantinos T. Konstantinidis[1,4*]

[1] Department of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

[2] Center for Environmental Diagnostics & Bioremediation, University of West Florida, 11000 University Parkway, Pensacola, FL, USA

[3] Department of Earth, Ocean and Atmospheric Sciences, Florida State University, Tallahassee, FL, USA

[4] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

[5] School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA

***Address correspondence to Konstantinos Konstantinidis, kostas@ce.gatech.edu**

Pages: 21

Supplemental Figures: 9

Supplemental Tables: 3

CONTENTS

S1: SUPPLEMENTAL METHODS

Community DNA and RNA extraction

Community DNA was extracted from the sand samples using the DNeasy PowerSoil kit (Qiagen Inc.) using the manufacturer's protocol. Community RNA was extracted using a modified protocol of the DNeasy PowerMax Soil Kit (Qiagen Inc.) designed for low biomass soil samples. The extracted RNA was treated with TURBO DNase, 2U/L (Invitrogen) and the RNA integrity verified on a Bioanalyzer 2100 instrument using an RNA Pico 6000 Kit (Agilent). Elimination of contaminating DNA from the RNA extracts were verified by performing PCR assays using the 16S rRNA gene using universal primers 8F and 1492R, followed by gel-electrophoresis. cDNA libraries were prepared using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Illumina, San Diego, CA, USA) and average insert sizes of the libraries determined to range from 361 bp to 437 bp using a DNA HS Kit (Agilent) . DNA and RNA concentrations were quantified using the Qubit 2.0 fluorometer with Qubit HS DNA and HS RNA kits (ThermoFisher Scientific), respectively. DNA libraries were prepared using the Illumina Nextera XT DNA library prep kit according to manufacturer's instructions except the protocol was terminated after isolation of cleaned double stranded libraries. Average insert sizes of DNA libraries and concentrations were determined as described above. An equimolar pool of the sequencing libraries was sequenced on an Illumina HiSeq 2500 instrument (School of Biological Sciences, Georgia Institute of Technology) using the HiSeq Rapid PE Cluster Kit v2 and HiSeq Rapid SBS Kit v2 (Illumina) for 300 cycles (2 x 150 bp paired end). Adapter trimming and demultiplexing of sequenced samples was carried out by the HiSeq2500 instrument.

Recovery of reads carrying 16S rRNA gene fragments from shotgun metagenomes

16S ribosomal rRNA (16S) gene fragments were extracted using Parallel-META 2.0 [1]and taxonomically classified using the RDP database[2]. Recovered 16S fragments were clustered ('closed-reference I picking' strategy) using UCLUST [3] and taxonomically classified using best match hits in the RDP database at an ID ≥ 97% in QIIME [4].

Alkane monooxygenase (*alkB*) gene diversity

Alkane monooxygenase (*alkB*) gene diversity was estimated based on the short reads, as described previously [5]. In brief, the *AlkB* reference protein sequences were aligned using CLUSTAL Omega [6] and a maximum likelihood reference tree was created using RAxML v 8.0.19 [7] based on a general time reversible model option with gamma parameter optimization and the '-f a' option. The identified *AlkB*-encoding reads were extracted from all datasets, translated into protein sequences using FragGeneScan, and then added to this reference alignment using Mafft [8]. The reads were placed in the phylogenetic tree using the RAxML EPA algorithm and visualized using iTOL [9].

Assembly and population genome binning

Co-assembly of the short reads from the biological triplicates was performed using IDBA-UD (Peng et al. 2012) and only resulting contigs longer than 500 bp were used for further analysis (e.g., functional annotation, binning and MyTaxa classification). MetaGeneMark [10] was used for gene prediction on the contigs. Contigs longer than 1Kbp were binned into individual population genomes or MAGs using an iterative binning technique to enable recovery of high

quality draft genomes [11]. The bins were checked for quality (contamination and completeness) using CheckM [12] and bins having a quality score of >50 (Completeness -5*Contamination) were used in the analyses. The intra-population diversity and sequence discreteness of the generated MAGs were determined using read recruitment plots that were generated with the scripts of the enveomics collection [13], as previously described [14].

Functional annotation of MAGs

Protein-coding regions of MAGs were functionally annotated based on best-match searches against the Swiss-Prot [15] database using a Blastp search. Only matches with a bitscore >60, amino acid identity ≥ 40% and alignment of 70% were used. The relative abundance of each function was subsequently estimated based on the number of predicted genes from each MAG assigned to that function and reported in Figure 7. Selected functions of interest such as biomarker genes for oil degradation were manually verified by examining the Blast output and/or the conservation of genes in the same operon.

Table S1: Summary statistics of the trimmed metagenomic and metatranscriptomic reads.

Sample	Condition	Trimmed Reads (MG)	Assembly N50 (bp)	Total Length (bp)	Trimmed Reads (MT)
T0	Control	28,889,704	1177	26,541,040	NA
T1O	Oil: Anoxic	53,097,416	1456	71,775,357	NA
T2C	Control	28,289,778	1100	22,459,221	95,084,232
T2O	Oil: Aerobic	44,876,158	2863	147,827,334	142,175,284
T3O	Oil: Anoxic	48,216,838	3177	84,843,636	137,521,458
T4C	Control	23,306,214	1059	14,578,253	106,087,022
T4O	Oil: Aerobic	46,297,004	2231	111,155,481	152,786,482

MG=Metagenome, MT=Metatranscriptome
The statistics shown above for each sampling time was the total of 3 biological replicates combined.

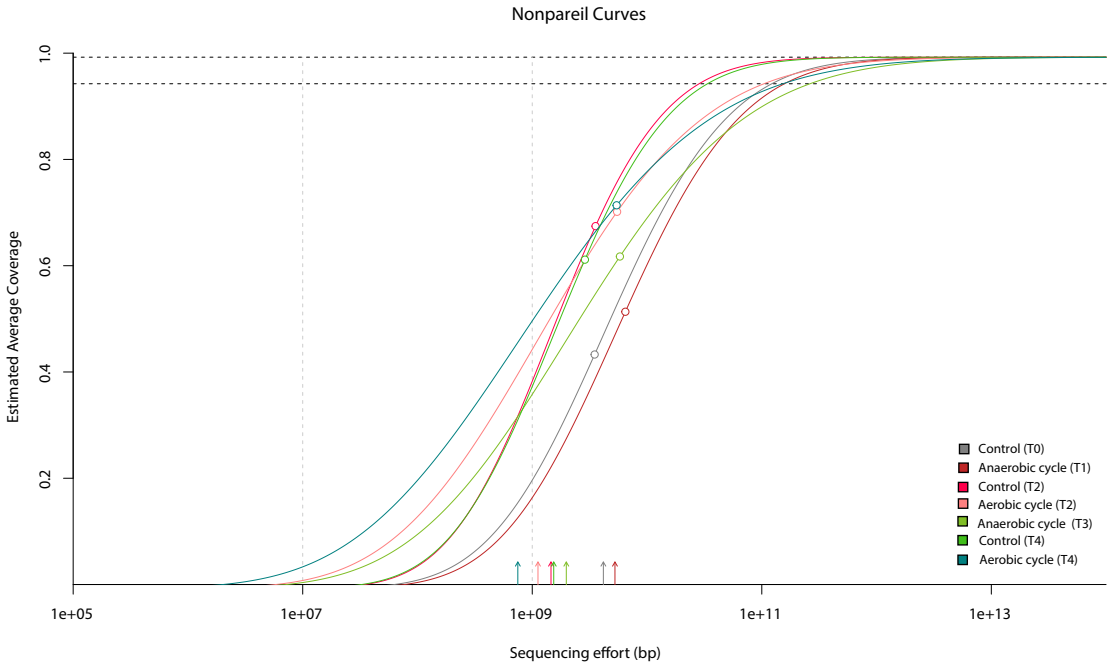
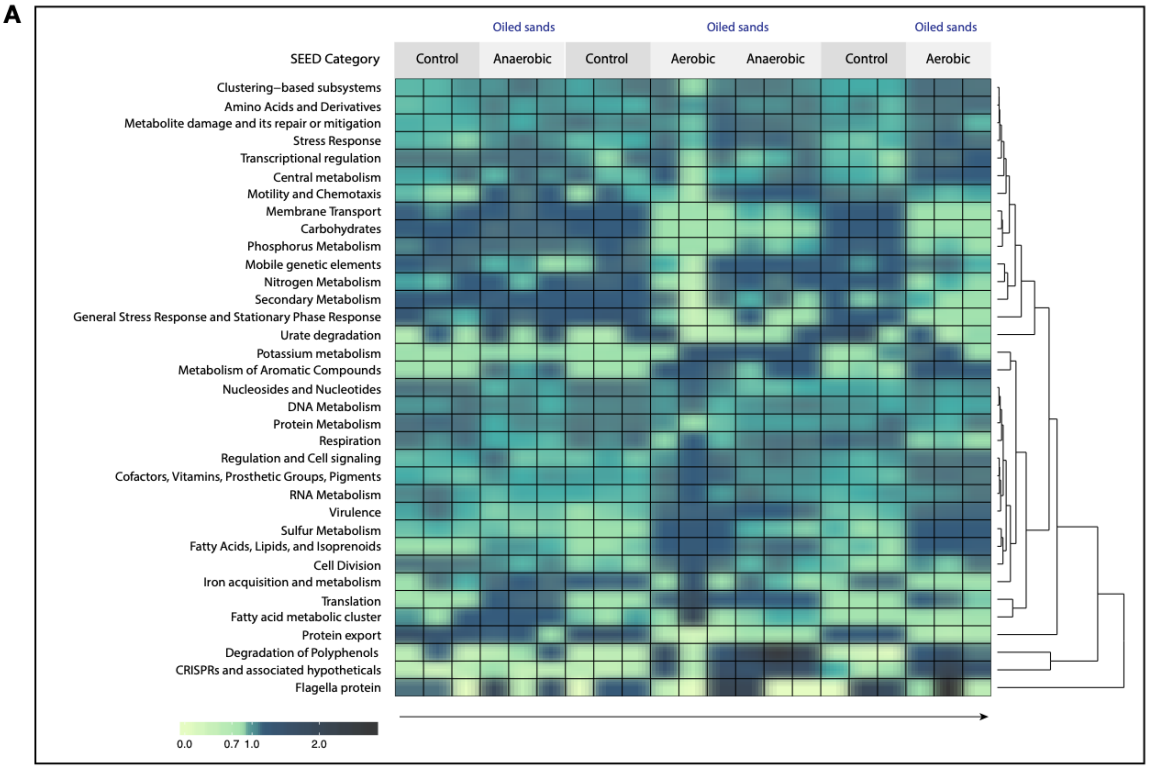


Figure S1: Nonpareil diversity and sequencing coverage estimates of the microbial communities sampled from mesocosm incubations. Empty circles represent the estimated average coverage of the datasets obtained and projections based on model fitting to reach 95% and 99% coverage are indicated (horizontal dashed lines. The arrows at the bottom represent sequencing effort required to achieve 50% coverage.


A
Oiled sands
Oiled sands
Oiled sands
SEED Category
Control
Anaerobic
Control
Aerobic
Anaerobic
Control
Aerobic
Clustering–based subsystems
Amino Acids and Derivatives
Metabolite damage and its repair or mitigation
Stress Response
Transcriptional regulation
Central metabolism
Motility and Chemotaxis
Membrane Transport
Carbohydrates
Phosphorus Metabolism
Mobile genetic elements
Nitrogen Metabolism
Secondary Metabolism
General Stress Response and Stationary Phase Response
Urate degradation
Potassium metabolism
Metabolism of Aromatic Compounds
Nucleosides and Nucleotides
DNA Metabolism
Protein Metabolism
Respiration
Regulation and Cell signaling
Cofactors, Vitamins, Prosthetic Groups, Pigments
RNA Metabolism
Virulence
Sulfur Metabolism
Fatty Acids, Lipids, and Isoprenoids
Cell Division
Iron acquisition and metabolism
Translation
Fatty acid metabolic cluster
Protein export
Degradation of Polyphenols
CRISPRs and associated hypotheticals
Flagella protein
0.0
0.7
1.0
2.0

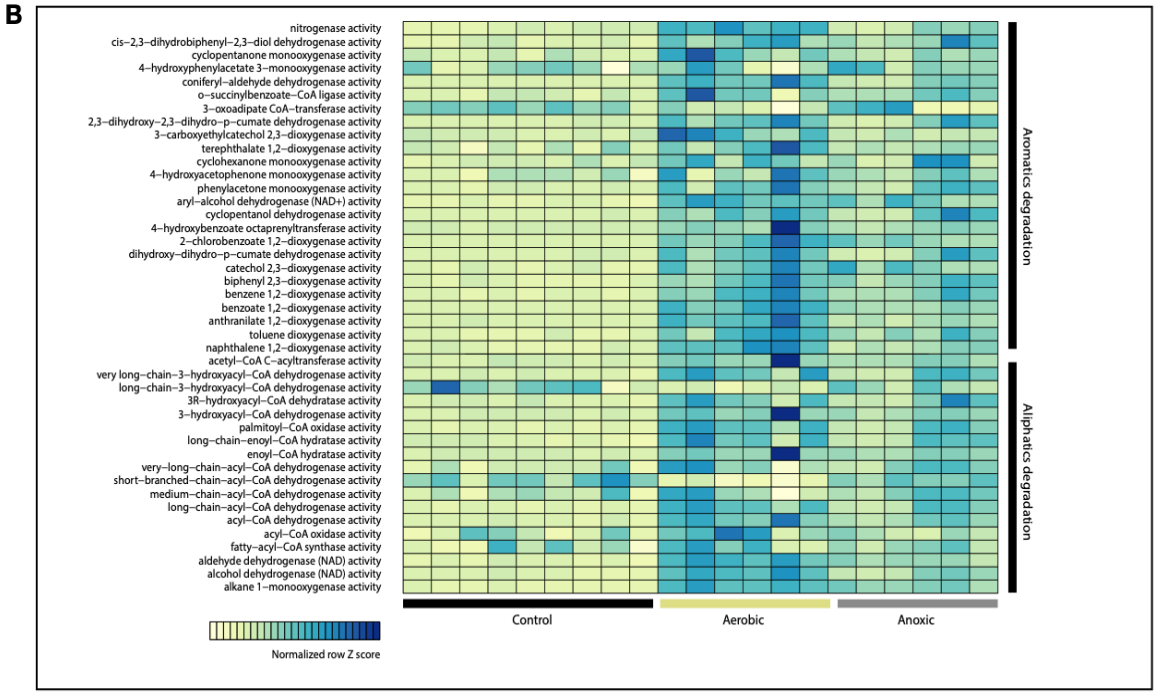

B
nitrogenase activity
cis–2,3–dihydrobiphenyl–2,3–diol dehydrogenase activity
cyclopentanone monooxygenase activity
4–hydroxyphenylacetate 3–monooxygenase activity
coniferyl–aldehyde dehydrogenase activity
o–succinylbenzoate–CoA ligase activity
3–oxoadipate CoA–transferase activity
2,3–dihydroxy–2,3–dihydro–p–cumate dehydrogenase activity
3–carboxyethylcatechol 2,3–dioxygenase activity
terephthalate 1,2–dioxygenase activity
cyclohexanone monooxygenase activity
4–hydroxyacetophenone monooxygenase activity
phenylacetone monooxygenase activity
aryl–alcohol dehydrogenase (NAD+) activity
cyclopentanol dehydrogenase activity
4–hydroxybenzoate octaprenyltransferase activity
2–chlorobenzoate 1,2–dioxygenase activity
dihydroxy–dihydro–p–cumate dehydrogenase activity
catechol 2,3–dioxygenase activity
biphenyl 2,3–dioxygenase activity
benzene 1,2–dioxygenase activity
benzoate 1,2–dioxygenase activity
anthranilate 1,2–dioxygenase activity
toluene dioxygenase activity
naphthalene 1,2–dioxygenase activity
acetyl–CoA C–acyltransferase activity
very long–chain–3–hydroxyacyl–CoA dehydrogenase activity
long–chain–3–hydroxyacyl–CoA dehydrogenase activity
3R–hydroxyacyl–CoA dehydratase activity
3–hydroxyacyl–CoA dehydrogenase activity
palmitoyl–CoA oxidase activity
long–chain–enoyl–CoA hydratase activity
enoyl–CoA hydratase activity
very–long–chain–acyl–CoA dehydrogenase activity
short–branched–chain–acyl–CoA dehydrogenase activity
medium–chain–acyl–CoA dehydrogenase activity
long–chain–acyl–CoA dehydrogenase activity
acyl–CoA dehydrogenase activity
acyl–CoA oxidase activity
fatty–acyl–CoA synthase activity
aldehyde dehydrogenase (NAD) activity
alcohol dehydrogenase (NAD) activity
alkane 1–monooxygenase activity
Aromatics degradation
Aliphatics degradation
Control
Aerobic
Anoxic
Normalized row Z score

Figure S2: Read-level functional gene shifts during mesocosm incubations. Heatmaps showing the relative abundance of major metabolic pathways (Level 1 of the SEED subsystems; Panel A) and individual gene functions related to oil biodegradation (Panel B). Note that in panel B, the samples are grouped according to "Treatment"; The 1st 3 columns under "Aerobic" refer to the 3 biological replicates during the first oxic phase and the next 3 represent the replicates during the 2nd aerobic phase.

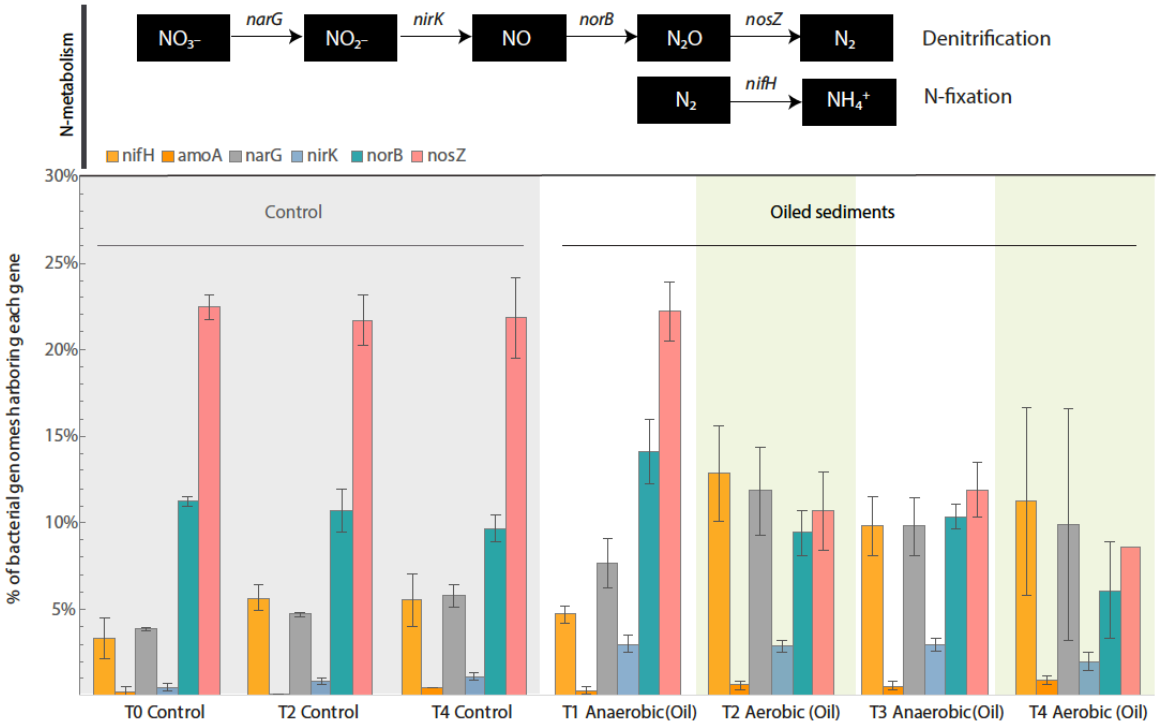


Figure S3: Abundance of nitrogen cycling genes and pathways over incubation time and treatment. Abundance of hallmark genes for denitrification, DNRA and nitrogen fixation pathways at each growth condition (see figure key) are shown, presented as genome equivalents (% of total genomes or cells encoding the gene, x-axis). The three samples per treatment represent biological replicates.

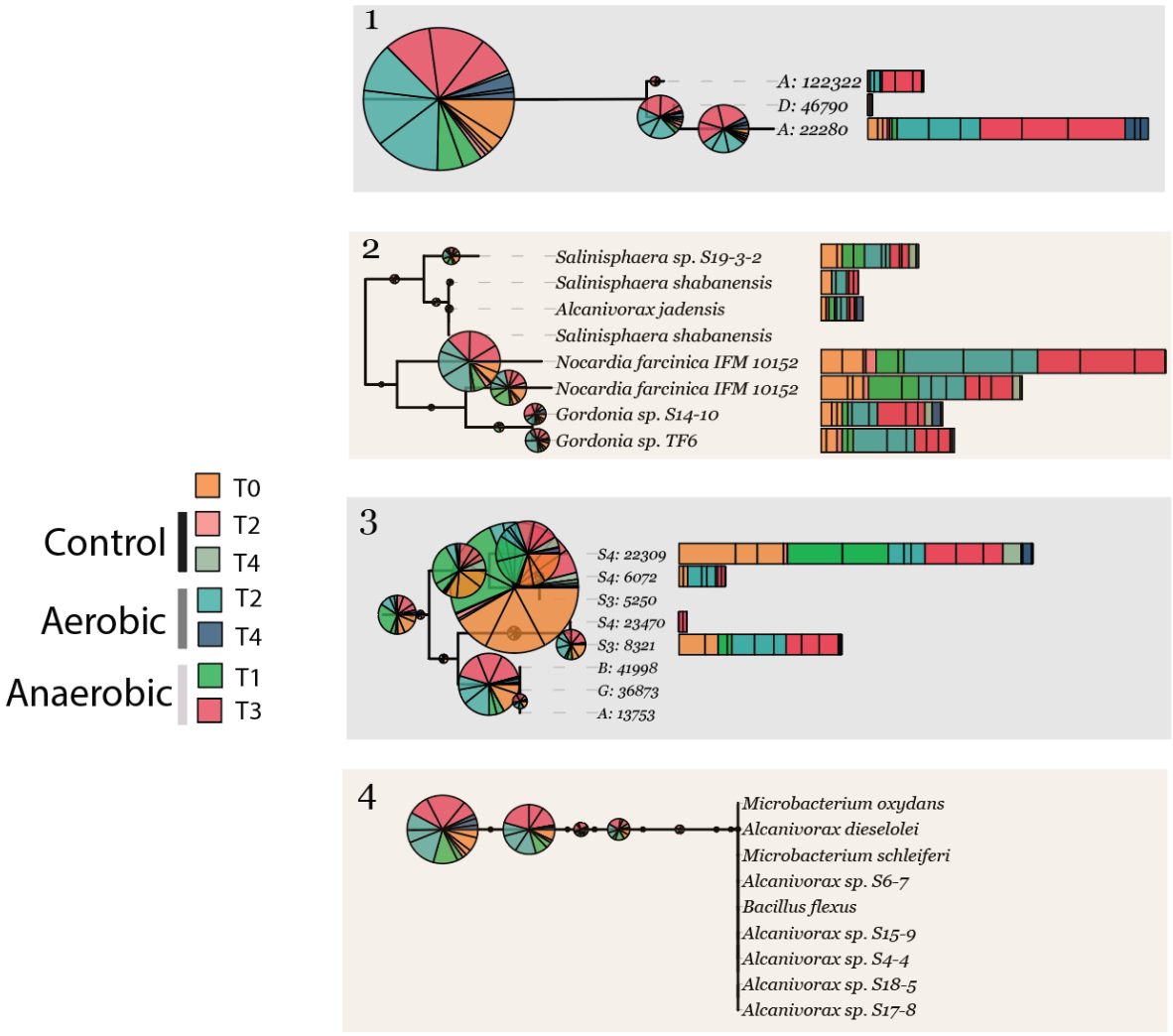


Figure S4: Phylogenetic diversity of *alkB*-containing metagenomic reads. Metagenomic reads assigned to the 4 sub-clades shown in Fig.3 of the main text. The radii of the pie charts and bars are proportional to the normalized number of reads assigned to a particular clade. Clades highlighted in gray indicate putative sequences recovered from the field datasets. Clades highlighted in red indicate putative sequences recovered from uncontaminated field sites..

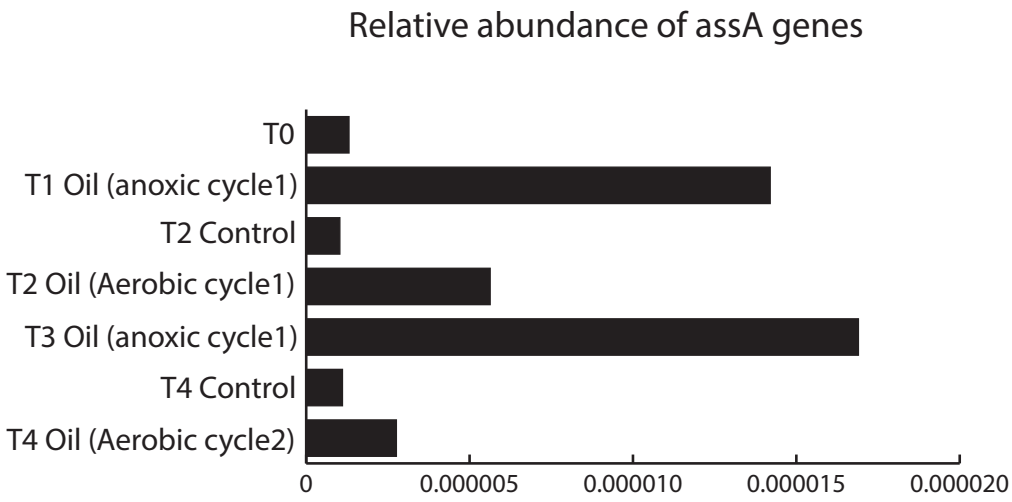


Figure S5: Relative abundance of the Alkyl/methylsuccinate synthase (*assA*) gene in the metagenomic datasets. Gene abundance is normalized by metagenome library size.

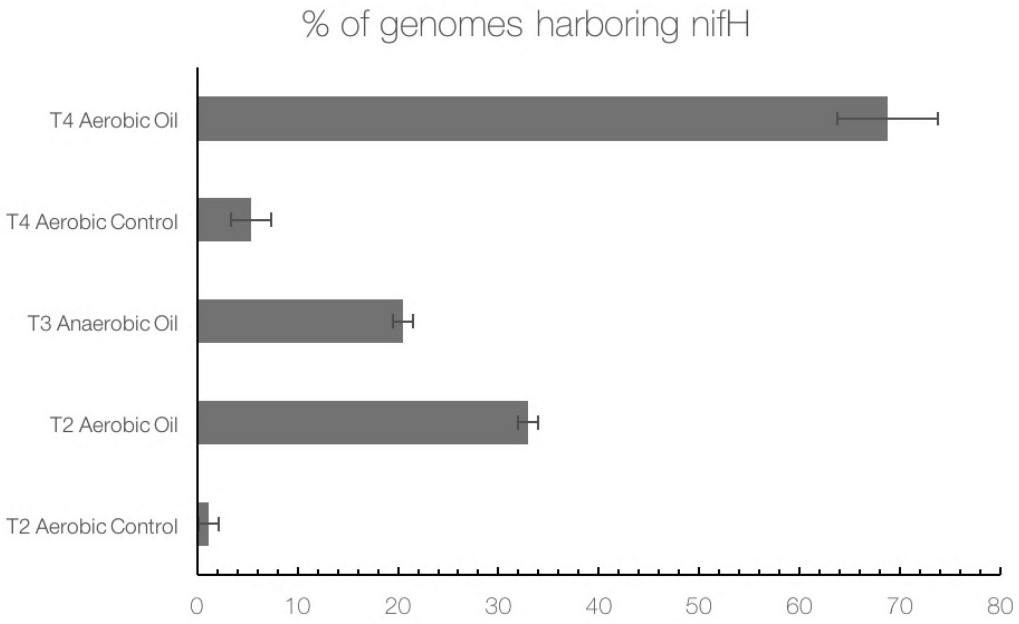
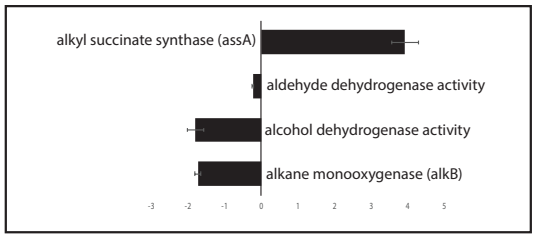
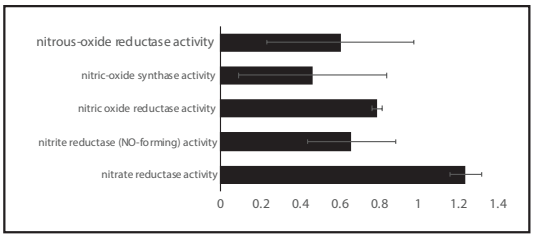


Figure S6: The fraction of genomes harboring the nitrogenase (*nifH*) gene through the time course of the study. Normalized gene abundance was calculated as genome equivalents (i.e. normalized by abundance of the single copy *rpoB* gene). The means of the 3 biological replicates for each condition and their SD are shown.

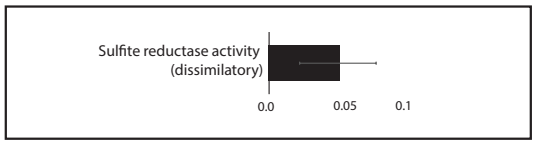
Alkane degradation



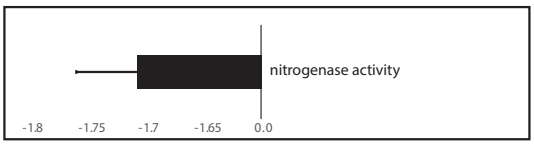
Denitrification



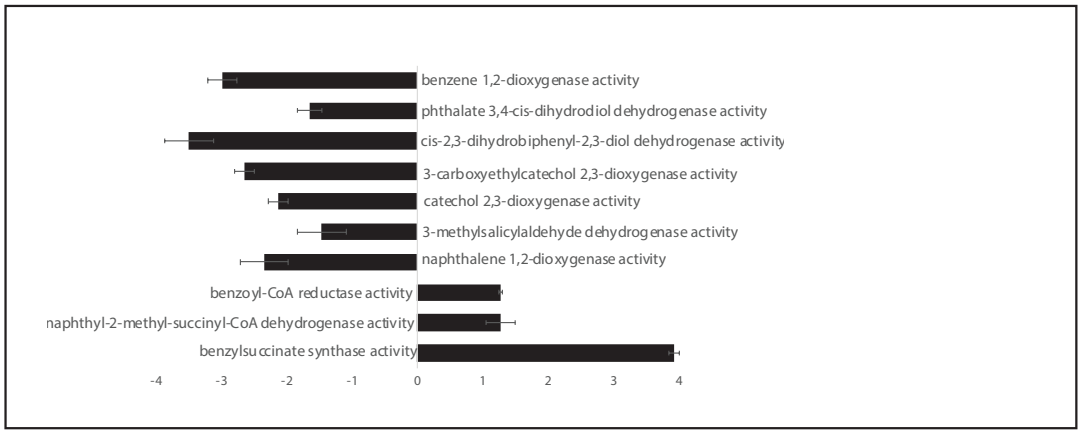
Sulfate reduction



N fixation



PAH/Aromatics degradation



More in aerobic ← → More in anoxic

Log 2-Fold Change

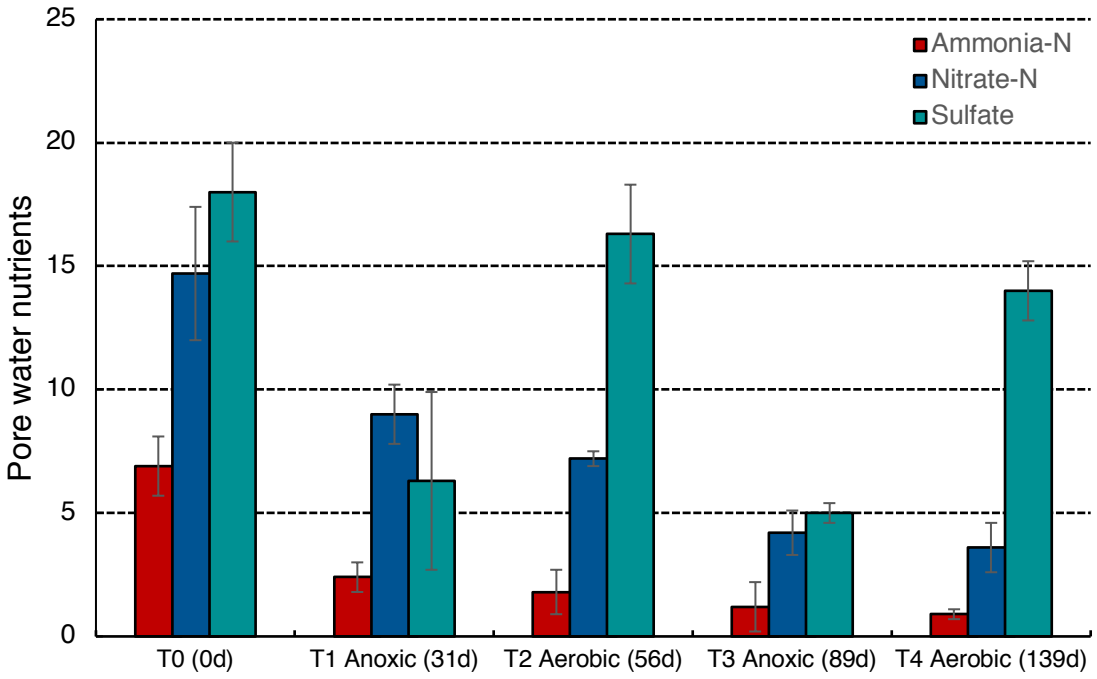


Figure S7: Upper panel: Log_2-fold changes in the expression of the associated pathway genes shown in Figure 5. Mean and SD of the three biological replicates are shown. **Lower panel: Measured pore-water nutrient concentrations in the oiled chambers.** Mean and SD of the three biological replicates through the run of the experiments are shown above. Note: Sulfate concentrations alone are in mM, rest are in μM. Nitrite-N concentrations were below detection in all chambers.

Figure S8: Successional patterns of MAGs and their gene-content during incubation. The solid squares (for aliphatic hydrocarbon degradation) and diamonds (for aromatic hydrocarbons) next to the genes indicate the presence of the respective gene (rows) in the MAG (columns). MAGs ordered along the x-axis are based on the sampling time point that they were recovered by genome binning. Note the increase in abundance of genes responsible for aromatics degradation over time (x-axis).

Table S2: Mean transcript abundance and log_2-fold change of functions associated with key metabolic pathways.

Function	BaseMean	log2FoldChange	pvalue	padj	SD
alkane monooxygenase activity	269.06171	-1.7286324	8.23E-27	3.90E-25	0.08063
alcohol dehydrogenase (acceptor) activity	31.31533	-1.8012773	5.15E-05	0.000225322	0.22245
aldehyde dehydrogenase (NAD) activity	3040.29054	-0.2286035	1.41E-05	6.74E-05	0.02632
alkylsuccinate synthase activity	15.60808	3.9226846	1.38E-07	9.02E-07	0.37232
benzylsuccinate synthase activity	15.60808	3.9226846	1.38E-07	9.02E-07	0.37232
naphthyl-2-methyl-succinyl-CoA dehydrogenase activity	67.38168	1.2851828	1.86E-05	8.65E-05	0.15011
benzoyl-CoA reductase activity	67.38168	1.2851828	1.86E-05	8.65E-05	0.15011
naphthalene 1,2-dioxygenase activity	12.74403	-2.3374009	0.001671502	0.00514494	0.37183
3-methylsalicylaldehyde dehydrogenase activity	49.96842	-1.4621001	0.000115616	0.000450091	0.18963
catechol 2,3-dioxygenase activity	36.86672	-2.1172907	1.97E-06	1.10E-05	0.22259
3-carboxyethylcatechol 2,3-dioxygenase activity	13.55905	-2.6316916	0.000545724	0.001848	0.38061
cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase activity	10.08017	-3.4872680	0.001147277	0.00369139	0.53623
phthalate 3,4-cis-dihydrodiol dehydrogenase activity	38.05732	-1.6364307	9.31E-05	0.000380002	0.20938
benzene 1,2-dioxygenase activity	10.08918	-2.9819520	0.001158675	0.00371616	0.45892
nitrogenase activity	846.58472	-1.7072857	4.75E-60	9.21E-58	0.05223
nitrate reductase activity	1049.63461	1.2437803	2.25E-09	1.85E-08	0.10402
nitrite reductase (cytochrome, ammonia-forming) activity	420.34751	-3.4551552	2.45E-88	1.04E-85	0.07332
nitrite reductase (NO-forming) activity	292.40338	0.6656003	5.65E-06	2.90E-05	0.07332
nitric oxide reductase activity	87.48158	0.7945402	0.008547091	0.05231455	0.15107
nitrous-oxide reductase activity	115.79020	0.6112885	0.005311202	0.03500082	0.10965
sulfite reductase activity	87.47903	0.0548118	0.821136673	0.930841286	0.12122
siderophore uptake transmembrane transporter activity	40.90580	-1.7672823	6.18E-06	3.13E-05	0.19549

Table S3: Taxonomic classification of the recovered MAGs.

Bin ID	*Closest Match in NCBI Prokaryote database*	AAI
T1 Oil Bins 001	*Alcanivorax sp N3 2A NZ CP022307*	67.8250814
T1 Oil Bins 002	*Thiohalobacter thiocyanaticus NZ AP018052*	48.443119
T1 Oil Bins 009	*Halomonas sp A3H3 NZ HG423343*	48.0112977
T1 Oil Bins 010	*Microbulbifer agarilyticus NZ CP019650*	51.4478783
T1 Oil Bins 011	*Desulfococcus oleovorans Hxd3 NC 009943*	47.7848973
T2 Oil Bins 002	*Thiohalobacter thiocyanaticus NZ AP018052*	48.1975444
T2 Oil Bins 004	*Candidatus Hodgkinia cicadicola CP024746*	40.7373262
T2 Oil Bins 005	*Hyphomonas sp Mor2 NZ CP017718*	56.0852036
T2 Oil Bins 007	*Thiohalobacter thiocyanaticus NZ AP018052*	46.157959
T2 Oil Bins 009	*Microbulbifer thermotolerans NZ CP014864*	51.3619975
T2 Oil Bins 010	*Candidatus Phaeomarinobacter ectocarpi NZ HG966617*	51.7179609
T2 Oil Bins 011	*Parvibaculum lavamentivorans DS 1 NC 009719*	50.4118297
T2 Oil Bins 012	*Thiohalobacter thiocyanaticus NZ AP018052*	47.9870392
T2 Oil Bins 014	*Candidatus Hodgkinia cicadicola CP024746*	42.5581624
T2 Oil Bins 015	*Sediminispirochaeta smaragdinae DSM 11293*	46.7222702
T2 Oil Bins 016	*Immundisolibacter cernigliae NZ CP014671*	46.6501682
T2 Oil Bins 019	*Immundisolibacter cernigliae NZ CP014671*	64.3240804
T2 Oil Bins 026	*Candidatus Hodgkinia cicadicola CP024746*	41.0779104
T2 Oil Bins 032	*Marinobacter sp CP1 NZ CP011929*	43.2009382
T2 Oil Bins 033	*Marinobacter hydrocarbonoclasticus ATCC 49840*	75.517838
T2 Oil Bins 034	*Thiohalobacter thiocyanaticus NZ AP018052*	45.18922
T2 Oil Bins 035	*Marinobacter hydrocarbonoclasticus ATCC 49840*	52.9382828
T2 Oil Bins 036	*Microbulbifer aggregans NZ CP014143*	45.2228105
T2C Bins 001	*Thiohalobacter thiocyanaticus NZ AP018052*	48.5854187
T2C Bins 002	*Maribacter sp HTCC2170 NC 014472*	60.5700871
T3 Oil Bins 001	*Thiohalobacter thiocyanaticus NZ AP018052*	48.1357197
T3 Oil Bins 002	*Desulfococcus_multivorans_NZ_CP015381*	48.1542287
T3 Oil Bins 003	*Thiohalobacter thiocyanaticus NZ AP018052*	46.2457598
T3 Oil Bins 004	*Parvibaculum lavamentivorans DS 1 NC 009719*	50.305553
T3 Oil Bins 005	*Alcanivorax sp NBRC 101098 NZ AP014613*	79.697554
T3 Oil Bins 006	*Desulfococcus_multivorans_NZ_CP015381*	43.61
T3 Oil Bins 010	*Thiohalobacter thiocyanaticus NZ AP018052*	48.0436986
T3 Oil Bins 011	*Micavibrio aeruginosavorus ARL 13 NC 016026*	49.7828113
T3 Oil Bins 012	*Desulfococcus_multivorans_NZ_CP015381*	51.2911452
T3 Oil Bins 016	*Thiohalobacter thiocyanaticus NZ AP018052*	48.8381799

T3 Oil Bins 018	*Desulfococcus oleovorans Hxd3 NC 009943*	47.7848973
T3 Oil Bins 019	*Muricauda lutaonensis NZ CP011071*	62.64011
T4 Oil Bins 001	*Kangiella geojedonensis NZ CP010975*	86.9806462
T4 Oil Bins 002	*Thiohalobacter thiocyanaticus NZ AP018052*	48.1571213
T4 Oil Bins 003	*Candidatus Campbellbacteria bacterium GW2011*	42.9542287
T4 Oil Bins 004	*Microbulbifer thermotolerans NZ CP014864*	50.9540735
T4 Oil Bins 005	*Hyphomonas sp Mor2 NZ CP017718*	55.728639
T4 Oil Bins 006	*Thiohalobacter thiocyanaticus NZ AP018052*	46.1026918
T4 Oil Bins 010	*Candidatus Phaeomarinobacter ectocarpi NZ HG966617*	49.7369432
T4 Oil Bins 011	*Immundisolibacter cernigliae NZ CP014671*	64.4579846
T4 Oil Bins 019	*Candidatus Hodgkinia cicadicola CP024746*	41.6172973
T4 Oil Bins 020	*gamma proteobacterium HdN1 NC 014366*	51.6325379
T4 Oil Bins 024	*Marinobacter salinus NZ CP017715*	60.7624307
T4 Oil Bins 025	*Muricauda ruestringensis DSM 13258 NC 015945*	48.238603
T4 Oil Bins 026	*Candidatus Hodgkinia cicadicola CP024746*	41.8003557
T4C Bins 001	*Zobellia galactanivorans NC 015844*	57.753619
T4C Bins 002	*Thiohalobacter thiocyanaticus NZ AP018052*	48.3222583

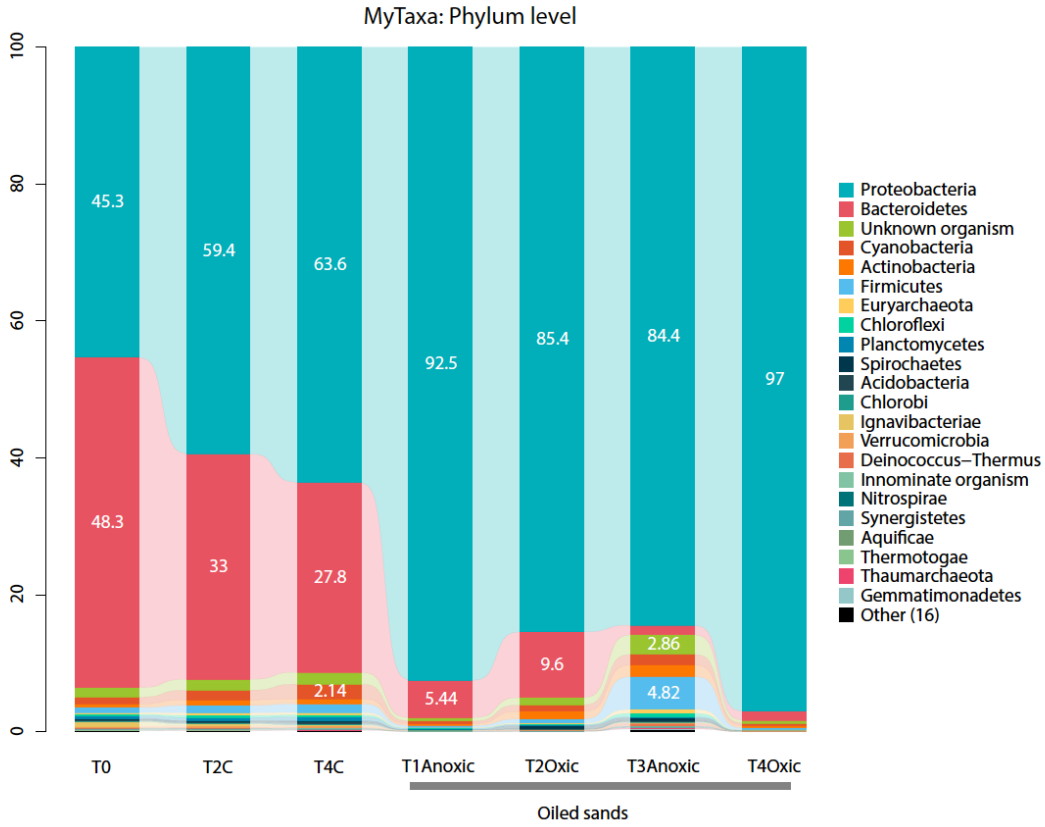


Figure S8: MyTaxa classification of the co-assembled reads from each sampling time point.

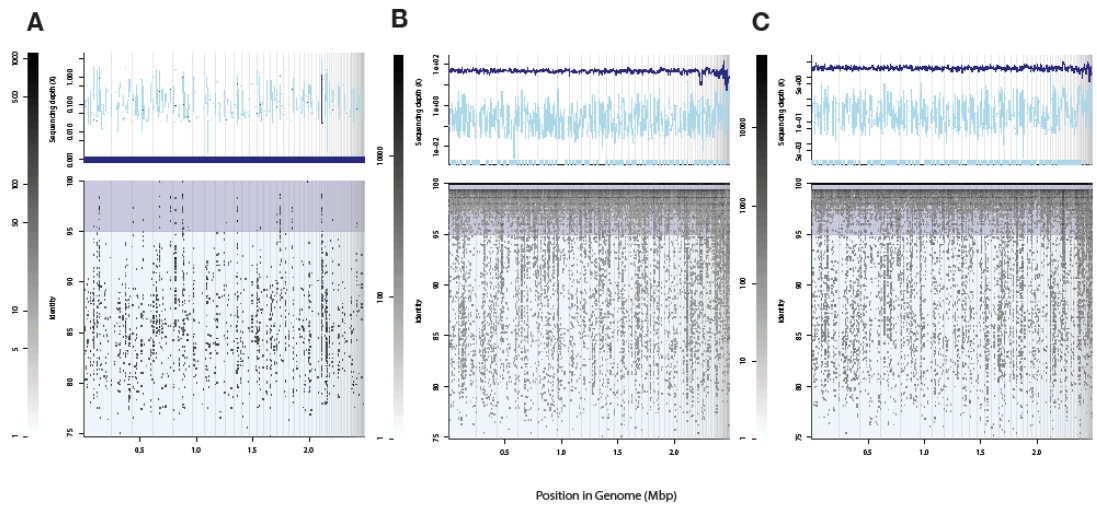


Figure S9: Read-recruitment using *Ca*. Macondimonas diazotrophica as the reference genome in control (A), T2 Aerobic (B) and T4 Aerobic (C) samples. The graphs represent the coverage of the *Ca*. Macondimonas diazotrophica genome sequence, in 1,000bp-long windows, by the metagenomic reads at three time points. The dark blue histogram in the lower panels represents the coverage by reads matching the reference genome at ≥80bp in length and ≥95% nucleotide identity; light blue represents reads matching at <95% identity. Note the two-fold change in abundance from T2 to T4 (dark blue line on upper panels). The even coverage across the genome reflects a homogenous, sequence-discrete population.

REFERENCES

1. Su, X.; Pan, W.; Song, B.; Xu, J.; Ning, K., Parallel-META 2.0: Enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization. *PLOS ONE* **2014,** *9*, (3), e89323.

2. Cole, J. R.; Wang, Q.; Fish, J. A.; Chai, B.; McGarrell, D. M.; Sun, Y.; Brown, C. T.; Porras-Alfaro, A.; Kuske, C. R.; Tiedje, J. M., Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research* **2014,** *42*, (Database issue), D633-D642.

3. Edgar, R. C., Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)* **2010,** *26*, (19), 2460-1.

4. Kuczynski, J.; Stombaugh, J.; Walters, W. A.; González, A.; Caporaso, J. G.; Knight, R., Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Bioinformatics* **2011,** *Chapter 10*, Unit10.7-10.7.

5. Rodriguez-R, L. M.; Overholt, W. A.; Hagan, C.; Huettel, M.; Kostka, J. E.; Konstantinidis, K. T., Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *The ISME Journal* **2015,** *9*, (9), 1928-1940.

6. Sievers, F.; Higgins, D. G., Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci* **2018,** *27*, (1), 135-145.

7. Stamatakis, A., RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* **2014,** *30*, (9), 1312-1313.

8. Katoh, K.; Misawa, K.; Kuma, K. i.; Miyata, T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **2002,** *30*, (14), 3059-3066.

9. Letunic, I.; Bork, P., Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research* **2016,** *44*, (W1), W242-W245.

10. Zhu, W.; Lomsadze, A.; Borodovsky, M., Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **2010,** *38*, (12), e132.

11. Tsementzi, D.; Rodriguez-R, L. M.; Ruiz-Perez, C. A.; Meziti, A.; Hatt, J. K.; Konstantinidis, K. T., Ecogenomic characterization of widespread, closely-related SAR11 clades of the freshwater genus "*Candidatus* Fonsibacter" and proposal of *Ca*. Fonsibacter lacus sp. nov. *Systematic and Applied Microbiology* **2019,** *42*, (4), 495-505.

12. 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* **2015,** *25*, (7), 1043-1055.

13. Rodriguez-R, L. M.; Konstantinidis, K. T. *The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes*; 2167-9843; PeerJ Preprints: 2016.

14. Konstantinidis, K. T.; DeLong, E. F., Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *The ISME Journal* **2008,** *2*, (10), 1052-1065.

15. Bairoch, A.; Apweiler, R., The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **2000,** *28*, (1), 45-48.