INTEGRATING OMICS APPROACHES TO PROVIDE A SYSTEMS-LEVEL VIEW OF MICROBIAL COMMUNITY RESPONSES IN BENTHIC ECOSYSTEMS AFFECTED BY THE DEEPWATER HORIZON OIL SPILL

A Dissertation
Presented to
The Academic Faculty

by

Smruthi Karthikeyan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
May 2020

INTEGRATING OMICS APPROACHES TO PROVIDE A SYSTEMS-LEVEL VIEW OF MICROBIAL COMMUNITY RESPONSES IN BENTHIC ECOSYSTEMS AFFECTED BY THE DEEPWATER HORIZON OIL SPILL

Approved by:

Dr. Konstantinos T. Konstantinidis, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Joel E. Kostka
School of Biological Sciences
Georgia Institute of Technology

Dr. Spyros G. Pavlostathis
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Thomas DiChristina
School of Biological Sciences
Georgia Institute of Technology

Dr. Jim C. Spain
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: December 16, 2019

To my grandparents

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisor Dr. Kostas Konstantinidis for his guidance and support through the last four years as well as Dr. Jim C. Spain for motivating me to pursue a field that was an entirely new avenue for me. I'm deeply grateful for them taking a chance on me and providing me with opportunities that greatly expanded my horizon of thinking. I couldn't have imagined coming this far in my academic career without their constant support and mentorship. I would also like to thank the collaborators of the various projects I worked on during the course of my PhD, specifically, Dr. Joel Kostka and Dr. Markus Huettel who were instrumental in shaping my thesis. I also would like to express my gratitude to my committee members Dr. Spyros Pavlostathis and Dr. Thomas DiChristina for their valuable input and feedback. I especially enjoyed the coursework I took under them. Having a committee with such diverse backgrounds greatly helped in broadening my view of analyzing a system through a truly multidisciplinary approach. I would like to thank the members of the Kostas lab, Spain lab and Kostka lab for their help and support on numerous occasions. Zuzi, Eric, Minjae, Coto and Miguel were of immense help when I first started out in a field so alien to me. I would also like to thank Dr. Janet Hatt who was of immense help in my wet-lab ventures. I'm forever grateful for all the friends I made along the way. I would be remiss if I did not thank my Atlanta family: Divya and Deepesh whose friendship and support made Atlanta truly my second home, Abi and Kevin whose close friendship and water cooler chit-chat was something I always looked forward to, Sebastian, for the much needed coffee breaks, and, Minjae for being a pillar of support through this journey. Finally, and most of all I would like to thank my family; especially

my husband, Shrivatsh who has been through it all with me (including my multiple “freak outs” and “existential crises”) and kept me grounded, my parents, Radhika and Karthik, my brother, Tarun, my grandparents for their unwavering support and also my extended family. Their constant love and support is what made this journey so meaningful.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AAI	Average Amino Acid Identity
ANI	Average Nucleotide Identity
DWH	Deepwater Horizon
DNRA	Dissimilatory Nitrate Reduction to Ammonium
GoM	Gulf of Mexico
HGT	Horizontal Gene Transfer
MAG	Metagenome Assembled Genome
OTU	Operational Taxonomic Unit
PAH	Polycyclic Aromatic Hydrocarbons
PCoA	Principal Component Analysis
TPH	Total Petroleum Hydrocarbons

SUMMARY

Hundreds of millions of liters of petroleum hydrocarbons enter the environment every year as a result of natural oil seeps or anthropogenic activities and accidents. Biodegradation mediated by a complex network of microorganisms and their interactions with their physicochemical environment ultimately dictate the fate of these hydrocarbons. Most of these interactions remain elusive due to the limitations of traditional, culture-based approaches to study microbial activities *in-situ*, but integration of culture-independent analyses can provide novel insights into this issue. The Deepwater Horizon (DWH) oil spill, one of the largest accidental oil spills in history, had devastating impacts on vast areas of the open ocean, deep sea and coastal ecosystems. The complexity in tracking the oil spill that occurred at 1500 m below the ocean surface, coupled to the dispersion of the oil due to the application of chemical dispersants, resulted in large amounts of oil being unaccounted for. Over 22000 metric tons of Macondo oil and surface residue balls (SRBs) or tarballs washed up on shore, severely impacting the coastal cities along the Gulf of Mexico (GoM). Microbial and geochemical hydrocarbon modifications during the degradation process resulted in two-fold higher complexity of the buried oil that reached the shores compared to the source well oil. The long term impacts of buried oil in these vulnerable ecosystems is still not well elucidated. The advent of next generation sequencing has opened new avenues of research thus enabling us to mine the "unseen majority" which typically comprise more than 99% uncultivated and not-previously described taxa, and thus, obtain new insights into the issues mentioned above and the complexity of microbial interactions during crude oil biodegradation.

Focusing primarily on coastal beach sand habitats, this thesis employed genome-resolved metagenomic approaches to isolate a novel, previously overlooked bacterial genus, *Candidatus* Macondimonas diazotrophica, that appeared to be a key member of coastal oil-biodegrading microbial communities (**Chapter 2**). A major limitation in predicting and modelling whole microbial community responses to oiling using "omics" techniques is the incorporation of flow dynamics/transport processes that can account for the transport of oxygen and nutrients to the buried oil, thereby affecting microbial activities. By integrating taxonomic, genetic and degradation rate data from laboratory advective flow chambers that simulate the beach sand environment, the responses of microbial communities to buried oil were elucidated **(Chapter 3**). Specifically, the microbial community successional patterns showed remarkable similarity to those obtained from the field data (sandy sediments affected by DWH) establishing robust and reliable biomarkers for evaluating the fate of buried oil in coastal sediments. Most notably, genes and taxa that were enriched during the initial phases were specialized for low molecular weight hydrocarbon degradation and where replaced by taxa specialized in high molecular weight PAHs at later degradation stages. Establishing reliable biomarkers to screen for oil degradation potential can also assist in determining whether an ecosystem is more "primed" for oil biodegradation or not **(Chapter 4)**.

During the past decade, millions of dollars were spent on genomic and metagenomic studies to understand how microbial communities responded to and bio-transformed the oil released by the DWH well in the GoM. Accordingly, public databases such as NCBI now host massive amounts of data in the form of metagenomes or metagenome assembled genomes (MAGs), but no database currently exists that integrates

these genomic/metagenomic data with physicochemical parameters known to affect the process of hydrocarbon biodegradation, or even documents the *in-situ* abundances of these genomes and their degree of association with oiled vs. clean samples. This gap in available resources severely limits data analysis and interpretation, and is not specific to oil but applies to additional processes of interest such as greenhouse gas emissions. To close this gap, we curated a database that includes microbial genomes (MAGs, single-cell amplified genomes or isolate genomes) binned from all publicly available metagenomic data associated with oil impacted environments (**Chapter 5**). The biogeographical distribution patterns of these genomes as well as their *in-situ* metadata are available as part of an interactive and searchable map of the database webserver (for details, see: http://microbial-genomes.org/projects/GROS). This repository should facilitate future studies to better understand the interactions among microbial community members and their chemical environment that ultimately control the fate of oil. The curated database can also serve as a model approach for building similar resources that combine omics data with physicochemical metadata for additional environmental processes and data of interest- a major bottleneck of several ongoing research and data integration efforts.

CHAPTER 1. INTRODUCTION

1.1 Marine Oil Spills: A global perspective

Hundreds of millions of liters of oil are released into the environment annually as a result of anthropogenic activities or natural hydrocarbon seeps (Head, Jones et al. 2006). Oil spills may occur due to accidental releases of crude oil from tankers, offshore platforms, pipelines, drilling rigs and wells, as well as spills of refined petroleum products and their by-products. All large oil spills to date namely, the *Ixtoc I* spill (1979), the *Amoco Cadiz* spill (1978), the *Exxon Valdez* oil spill (1989) and the Deepwater Horizon spill (2010), have affected marine ecosystems. More recently, studies have reported large amounts of oil being discharged from the Taylor Energy platform in the GoM as a result of destruction of the oil platform during Hurricane Ivan. The 14-year long spill is estimated to surpass the oil released from the Deepwater Horizon accident by an order of magnitude, making it the largest accidental oil spill in history (Sun et al. 2018). GoM alone, the site of the Deepwater Horizon, has over 3500 drilling platforms and is characterized by a plethora of natural oil and methane seeps whose oil seepage rates are estimated at over tens of millions of liters per year (Orcutt, Joye et al. 2010, Teske 2019). Oil spills are one of the most serious threats to marine ecosystems, and the long-term consequences of spills remain poorly understood. Several studies have shown the oil's detrimental impacts to coral communities, acute immunotoxic effects to the aquatic organisms as well as the mutagenic effects induced on phytoplankton communities (Barron 2011; White et al. 2012; Paul et al. 2013) Furthermore, the chemical dispersant, Corexit 9500, which was applied to disperse the oil (several million liters applied), has been shown to suppress microbial growth and

activity (Kleindienst, Seidel, et al. 2015). Coastal ecosystems are amongst the most vulnerable ecosystems to oil spills also because oil that reaches the shoreline can remain buried and thus may be preserved from erosion and bio- or photo-degradation for long periods of time. Investigating and better modeling of the fate of buried oil is thus essential for risk assessment of environmental and human health, coastal food sources, and economic impacts.

1.2 The Deepwater Horizon blowout, April 2010: Tracking the oil plume

The DWH oil spill occurred 40 miles off the coast of Louisiana, USA when a failure in the blowout preventor valve (BOP) resulted in what would become one of the largest environmental disasters in history. Over 780 million liters of oil and large amounts of natural gas were released into the GoM, resulting in widespread impacts on the benthic and coastal ecosystems of the Gulf (Atlas and Hazen 2011, King, Kostka et al. 2015). The oil leaked from the riser pipe of the sinking drilling rig at 1500 m below the sea level for about 84 days before it was finally capped. The DWH oil spill was the largest accidental marine oil spill in history and its scale and source depth (1500 m below sea level) represented a relatively uninvestigated category of oil spill and challenging to manage (Camilli, Reddy et al. 2010). The complexity of modelling the deep-sea plume dynamics coupled to the large amounts of natural gas that were also released, made it difficult to track the transport and fate of the oil. Consequently, about 22% of the spilled oil is still unaccounted for, which raises significant concerns about its environmental impacts (Mason et al. 2014; Bagby et al. 2017). It was estimated that approximately one-half of the oil released from the DWH blowout reached the ocean surface, and a portion of it was transported to nearshore and coastal ecosystems and buried in the sediments. Large amounts of weathered

oil washed up onshore across four states along the GoM covering an estimated 1773 km of the shoreline following the blowout. In order to defray the environmental and safety impacts of the leaking oil from the Macondo Prospect, the dispersant, COREXIT 9500 was applied directly at the wellhead or end of the riser pipe at a depth of 1500 m below the sea. The main goal of the addition of a dispersant was to increase the "bioavailability" of oil by helping overcome the surface tension at the oil-water interface, thereby making the oil more accessible for the microorganisms that have the potential to degrade the hydrocarbons. However, the efficacy of the applied dispersant in mitigating the spill still remains debated as negative effects on microbial activity from exposure to the dispersant have been noted (Kleindienst, Seidel, et al. 2015).

1.3 Factors governing the ultimate fate of released crude oil

Biodegradation mediated by a complex network of microorganisms controls the ultimate fate of the majority of oil hydrocarbons that enter the marine environment (Leahy and Colwell 1990). However, the interactions among microbes and with their chemical environment during oil biodegradation are highly complex and poorly understood. Figure 1-1-1 shows a simplified schematic of the microbial interactions involved in the biodegradation of crude oil components. Crude oil compositions are highly complex and the relative proportions of the hydrocarbons are highly variable, depending on the reservoir source and its underlying biogeochemistry (Overton et al. 2016). However, once the crude oil is released into to the marine environment, the components undergo rapid changes, which are highly dependent upon the *in-situ* physicochemical conditions. Depending on the size and chemical composition, the crude oil fractions are amenable to physicochemical processes like dissolution, evaporation etc. or biological processes like partial or complete

degradation by indigenous microbes. Some of the more recalcitrant fractions of oil such as the polycyclic aromatic hydrocarbons (PAHs) and asphaltenes tend to bioaccumulate or persist for prolonged periods of time. A significant amount of oil that reached the shorelines along the GoM had a much altered chemical composition compared to the crude oil at the seafloor due the extensive weathering and biological processes (Huettel et al. 2018).

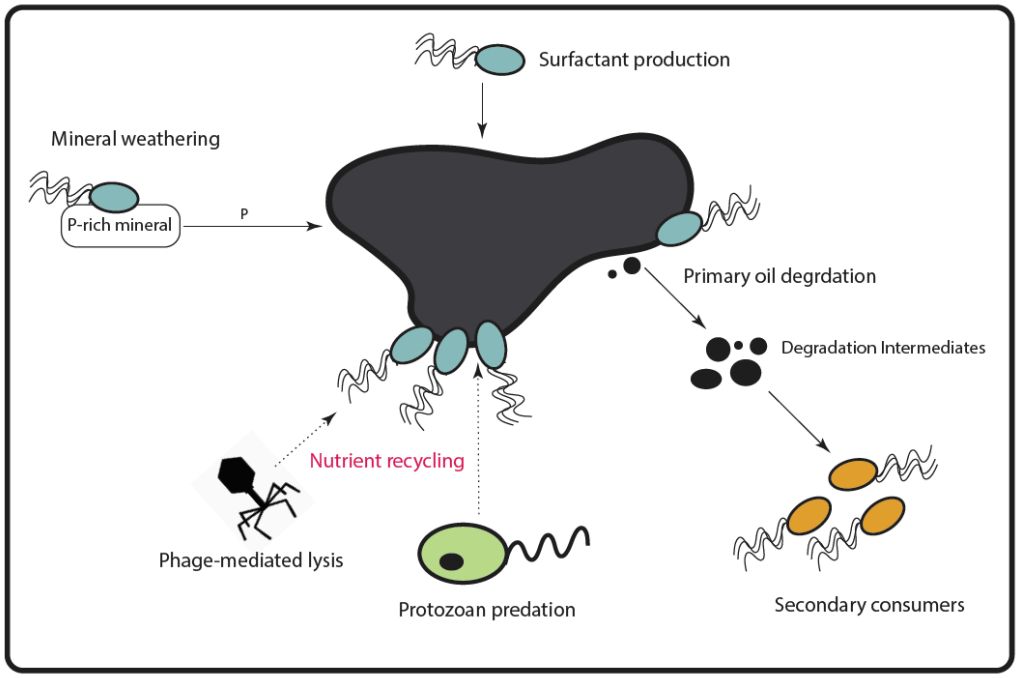


Figure 1-1: Schematic of the microbial oil biodegradation network. Solid arrows depict material fluxes, broken arrows indicate direct interactions (for instance, phage-mediated lysis and predation by protozoa). Interactions between primary oil degraders and their environment are also shown. Figure adapted from Head et al., 2006.

Most components of crude-oil are essentially biodegradable under oxic conditions albeit some of the heavier, recalcitrant compounds like hopanes, resins, asphaltenes can

have extremely slow rates of biodegradation (Prince and Atlas 2018; Atlas and Hazen 2011; Leahy and Colwell 1990; Gibson and Parales 2000). Anaerobic degradation of hydrocarbons with alternative electron acceptors have also been shown to be prevalent and widespread, although this process occurs, in general, at much slower rates compared to aerobic degradation (Widdel and Rabus 2001a; Shin et al. 2019; Kimes et al. 2013). Limitations of nutrients, including molecular oxygen and bioavailable nitrogen, is one of the major barriers that dictate the rate of degradation of hydrocarbons in the environment. Crude oil typically has low nitrogen content (<1%) thereby necessitating exogenous nitrogen inputs to stimulate biodegradation. *In-situ* biodegradation rates of crude oil are significantly dependent upon the concentration of rate limiting nutrients such as fixed nitrogen, phosphorous and iron (Atlas and Bartha 1972). For instance, shoreline clean-up operations in the wake of the Exxon Valdez oil spill in 1989 in Prince William Sound, Alaska involved the large-scale application of nitrogen based fertilizers. Over 48,600 kg of nitrogen in the form of fertilizers was applied across 1400 sites in Alaska over a period of two years (Pritchard et al. 1992). However, addition of large amounts of fertilizer (biostimulation) can also lead to an increase in harmful algal blooms as well as long-term alterations of the ecosystem community dynamics. Increased levels of fixed nitrogen input through a diazotrophic community could circumvent the need for fertilization in the event of an oil spill and the environmental consequences from the application of the fertilizer, albeit with reduction in the biodegradation rates.

1.4 Omics-based studies of oil degradation and ecosystem impacts

After coastal oil spills, petroleum hydrocarbons can accumulate in submerged nearshore sediments and on beaches, creating health risks for coastal organisms and

humans. The highly variable conditions of beach environments make it challenging to determine the long-term behavior and fate of hydrocarbons in beach sands and sediment. The DWH accident exposed the dearth of microbiological research on the GoM shores and the lack of baseline environmental data and models that could be used to formulate effective responses to an environmental disaster of this magnitude. While most studies to date have focused on the fate of oil in the deep sea plume and sediments, its impact on beach sand communities remains essentially understudied. This is due, at least in part, to the stochasticity and complexity of ecosystem processes and lack of appropriate model microorganisms. Addressing these knowledge gaps will lead to the elucidation of robust biomarkers of oil biodegradation, and thereby, an improved understanding of the fate of oil in nearshore ecosystems. Biomarkers that are representative of the different phases/stages in oil degradation (for instance, early, mid vs. late) and/or the limitations of microbial oil biodegradation (e.g., nitrogen shortage), possibly including oil-sensitive organisms, can help site managers to monitoring oil degradation and decide on the appropriate actions to biostimulate the biodegradation process (e.g., alter environmental conditions), with obvious benefits for restoring ecosystem functions and minimizing public and environment health risk.

1.4.1 Elucidating previously uncharacterized microbial taxa controlling oil biodegradation

Our previous study (Rodriguez-R et al. 2015) was one of a first efforts to characterize microbial community response to oil perturbation in the sandy intertidal zones (Pensacola Municipal Beach, FL). Beach sands were collected at Pensacola Municipal Beach, FL, USA before the arrival of the oil plume, one month after the oil reached the

beach, three months later when oil constituents were still present in the sand, and a year after, when oil was not detectable. Samples were collected from aerobic beach sediments down to 55cm depth. 16 shotgun metagenome datasets and 122 16S rRNA (16S) gene amplicon datasets were generated from these samples. The study showed, contrary to expectations, that generalist taxa (characterized by larger genome sizes and a high(er) number of catabolic pathways), associated primarily with *Gammaproteobacteria*, as opposed to specialists, were favored by the oil perturbation. This study also identified which individual populations responded to oil and their successional patterns until a year later, when oil was not detectable in the sandy sediments sampled. Furthermore, it revealed that greater than 95% of the responding taxa belonged to uncultured taxa, underscoring the latent untapped microbial diversity that exists in these ecological niches. In particular, early responders to oil contamination, likely degrading aliphatic hydrocarbons, were later replaced by populations capable of aromatic hydrocarbon decomposition, most of which represented novel taxa. Among the latter, a phylogenetically narrow group of oil degraders became highly abundant in oiled sediments that were below detection in uncontaminated sediments.

Following up on the results of this first study by Rodriguez-R et al, 2015, in **Chapter 2**, we employed a combination of genome-resolved metagenomic analyses as well as PCR walking in order to recover the genome and the 16S rRNA gene sequence of this key population that was highly abundant in oiled sediments of coastal marine ecosystems. Functional gene annotations revealed that this population possesses the genes required for hydrocarbon degradation as well as nitrogen fixation. This was a potentially important finding given that oil-degradation kinetics in these niches are often nitrogen limited. Efforts

to isolate representatives of these abundant populations, using the recovered genome from the metagenome as a guide, were successful, yielding several isolates (Karthikeyan et al. 2019). One of the isolates was fully sequenced and its genome encoded functional nitrogen fixation and hydrocarbon degradation genes together with putative genes for biosurfactant production that apparently facilitate growth in the typically nitrogen-limited, oiled environment. Comparisons to available genomes revealed that this isolate represented a novel genus within the *Gammaproteobacteria*, for which the provisional name **"*Candidatus* Macondimonas diazotrophica" gen. nov., sp. nov.,** was proposed (Karthikeyan et al. 2019). "*Ca.* M. diazotrophica" appears to play a key ecological role in the response to oil spills around the globe based on its high abundance in oiled but not clean or pristine coastal sediments and sandy sediments, often comprising ~30% of the total oiled microbial community, and thus, could be a promising model organism for studying ecophysiological responses to oil spills. Furthermore, the genes recovered in the "*Ca*. M. diazotrophica" genome probably account for its high ecological success in oiled samples and "*Ca*. M. diazotrophica" could be potentially used for bioaugmentation efforts in place of adding copious amounts of exogenous chemical nitrogen compounds during oil bioremediation projects.

1.4.2 Improving predictions of oil biodegradation and ecosystem recovery

Erosion and deposition cycles lead to the burial of weathered hydrocarbons in water saturated (subtidal), temporally saturated (intertidal) and unsaturated (supratidal) sands. Covered by anoxic sediment, oil may persist in largely un-weathered form and thus, contain relatively large concentrations of harmful PAHs that can be released from the sediments during storms. Deciphering crude oil biodegradation dynamics in anoxic sediments

represents a major challenge in managing oil spills. Prediction of the effects and fate of these crude oil products remains hampered by our limited understanding of the controls of biodegradation and functioning of sedimentary microbial communities. In order to better predict the coastal ecosystem response to such spills, models that link microbial community structuring and activities with the *in-situ* transport processes are essential. Although certain microbial community successional patterns have been established in response to hydrocarbon availability, we still have a very limited understanding of the universal applicability of these patterns, and how they may differ depending on the exact environmental settings and transport processes. It is still unclear, especially from a quantitative perspective, how these interactions affect the rate of ecosystem function/recovery and how the interactions may change under different *in-situ* conditions. The effect of regular oxygen oscillations in the intertidal zone and subtidal sediments on the microbially mediated oil degradation is presumably significant but challenging to quantify and model (Huettel et al. 2018). Whereas the flow of air through unsaturated beach sands can rapidly transport oxygen to buried oil, it cannot carry nutrients that are limiting the degradation of the oil. On the other hand, transport via pore water flows in submerged sands is slower than gas transport in dry sand, but water can transport dissolved nutrients to buried hydrocarbons. Quantifying the effects of such conditions on oil degradation is important to enable a more system-level view of microbial oil biodegradation in coastal sediments.

In **Chapter 3,** we employed advective flow laboratory chambers that have been previously shown to simulate the submerged benthic coastal environments to provide new insights into the oil biodegradation dynamics under *in-situ* conditions. A combination of

metatranscriptome and genome-resolved metagenomics was used to elucidate the controls of microbial oil degradation undergoing temporal, oxic-anoxic cycles in the laboratory incubations. Hydrocarbon quantification and metatranscriptomics analyses showed that oil biodegradation was not severely limited in the absence of oxygen, with sulfate and to a lesser extent, nitrate, serving as alternative electron acceptors in the anoxic phases. Metabolic pathway reconstruction and genome-resolved metagenomics provided key insights into the pathways employed by the microbial communities during the oil biodegradation phase. Interestingly, microbial activities during the oxic phases further promoted the anaerobic biodegradation by re-oxidizing (and/or detoxifying) the (reduced) alternative electron acceptors and providing nitrogen, an oil-limiting nutrient, through biological nitrogen fixation. Furthermore, comparison of the microbial community responses documented in these laboratory incubations closely mirrored pattern observed based on similar field data (Pensacola Beach, Florida, USA) (Rodriguez-R et al. 2015), revealing potentially reliable biomarkers to study these vulnerable ecosystems.

1.4.3 Establishing biomarkers of oil degradation in marine ecosystems

1.4.3.1 Alkane degradation genes

Alkanes are widespread and ubiquitous in nature, either as a result of anthropogenic activity or through biogenic inputs. As a result of the ecological pervasiveness of alkanes, indigenous microbes with the capacity to degrade them are widespread as well. In marine systems, widespread biogenic alkane production by algae has been documented. Marine algae/cyanobacteria produce mid- length straight-chain hydrocarbons (primarily C15 and C17 alkanes) to the tune of ~308–771 million tons annually (Schirmer, Rude et al. 2010,

Lea-Smith, Biller et al. 2015, Valentine and Reddy 2015). This estimate of biogenic alkane abundance in marine systems far exceeds the input from oil spills and natural seeps (White, Marx et al. 2019). (Schirmer, Rude et al. 2010). Due to the ubiquitous presence of cyanobacteria in surface marine waters, it has been hypothesized that a "latent" hydrocarbon cycle exists in the ocean where these algae-based hydrocarbons sustain the growth of obligate or facultative hydrocarbon degrading microbes thus priming these organisms to handle oil spills (Lea-Smith, Biller et al. 2015). However, crude oil is a complex mixture of hydrocarbons (including aromatic), and thus, oil-degrading bacteria must also break down complex types of hydrocarbons in oil that are not found among algal-produced hydrocarbons.

Alkanes are known to serve as chemo-attractants to several lineages of bacteria and their active transport mechanisms are hypothesized to determine the chain length that the bacteria have the capacity to degrade (Shao and Wang 2013). Accordingly, the enzymes responsible for the degradation of alkanes have been broadly classified into short-,mid-, and long-chain alkane degraders. The most frequently encountered ones belong to a class of integral membrane non-heme iron monooxygenases, the alkane monooxygenase or *alkB* which catalyzes the initial hydroxylation of mid-chain alkanes (~C17) (Rojo 2009). The strains carrying this enzyme also frequently carry the soluble cytochrome P450 enzymes. Other enzymes of the hydroxylase family implicated in degrading longer chain alkanes are *AlmA* (Throne-Holst et al. 2007), a putative monooxygenase belonging to the flavin-binding family, and, *LadA,* an oxygenase belonging to the luciferase family (Li et al. 2008). *AlmA* and *LadA* share no apparent homology with the *alkB* or cyt-P450 enzymes. However, to date the genes encoding *alkB* are the most studied due to their ubiquity as well as the

availability of experimental evidence for their function. *alkB* and cyt-P450 gene expression if tightly regulated and the transcriptional regulators are frequently found in close proximity to *alkB* and cyt-P450 genes but the details of the regulation and how different alkanes affect gene expression have not been elucidated yet (Shao and Wang 2013).Incorporation of "omics" based techniques can help further our understanding of the global metabolic networks as well as the overall process of bacterial alkane-dependent chemotaxis, alkane transport, gene expression regulation and complete mineralization. It remains currently unclear whether oil-degrading microbes use the same *AlkB* genes for degradation of crude oil as for cyanobacterial-produced hydrocarbons or, instead, have evolved specialized *AlkB* genes for crude-oil degradation favor (selected by) the presence of natural oil-seeps.

Genes encoding *AlkB* (alkane hydrolase) are generally considered as a biomarker for alkane degradation. In spite of its ecological pervasiveness, the reliability of *alkB* as a potential biomarker for crude-oil biodegradation potential has not been established yet. Most of the earlier studies describing *alkB* sequences and their phylogenetic distribution across clades were based upon "known" or isolate sequences or targeted amplicon PCR efforts (Kloos, Munch et al. 2006, Wasmund, Burns et al. 2009, Wang, Wang et al. 2010, Shao and Wang 2013, Smith, Tolar et al. 2013, Nie, Chi et al. 2014). Yet, recent work has shown that the great majority of the populations responding to crude oil (>95% of the total) represent novel taxa not closely related to the isolated and/or characterized ones ((Karthikeyan, Rodriguez-R, Heritier-Robbins, Hatt, et al. 2019); discussed also above). In **Chapter 4**, a large scale analysis of (meta-) genomic data revealed a novel, divergent *alkB* clade with no cultured/described representatives that was recovered exclusively from

crude-oil impacted ecosystems including those affected by the Deepwater Horizon oil spill. By contrast, the *alkB* clades associated with consumption of biogenic alkanes (cyanobacterial associated) clades belonged to "canonical" or hydrocarbonoclastic clades.

1.4.3.2 Is the Gulf of Mexico "primed" to handle oil-spills?

The site of the DWH incident, the northern GoM (nGoM), is characterized by plenty of natural oil and methane seeps whose oil seepage rates are estimated at over 20,000 m^3/yr (Orcutt, Joye et al. 2010, Kleindienst, Grim et al. 2015, Teske 2019). This, coupled to the relatively lighter nature of the spilled crude-oil (thus, more inherently amenable to biodegradation) was viewed as the major reason for the rapid consumption of a significant portion of the released oil (Hazen, Dubinsky et al. 2010, Atlas and Hazen 2011). Following the DWH blowout, specific taxonomic shifts were observed across spatial and temporal scales that were characterized by a short-term decrease in the microbial community diversity. (Mason, Hazen et al. 2012, Kimes, Callaghan et al. 2014, Lamendella, Strutt et al. 2014, Rodriguez-R, Overholt et al. 2015). Interestingly, the most predominant responders across all impacted ecosystems belonged to uncultured taxa rather than the well described hydrocarbonoclastic genera. Previous oligotyping based studies have also indicated that diverse taxa from the rare biosphere, which likely represent sequence discrete, distinct ecotypes, shaped the deep sea plume associated microbial community (Kleindienst, Grim et al. 2015). However, whether these "key" responders are always members of the rare biosphere in the clean (not oiled) GoM sites and in other oceans globally has not been quantified. Quantitively testing whether or not the abundances of these responding species are higher in the clean samples of the GoM relative to those from

similar habitats across the world can provide novel insights into the degree to which the GoM may be "primed" to deal with sudden anthropogenic oil inputs or spills.

In chapter 5, we screened over 265 metagenomes spanning the world's oceans for the presence of the abundant populations recovered from the DWH-impacted samples. The analysis revealed the presence of over 40% of these MAGs in the uncontaminated GoM water column samples. By contrast, the global marine metagenome data (exclusively uncontaminated samples from the TARA Oceans expedition) contained (similar relative abundances) only the genomes of known hydrocarbonoclastic bacteria namely, *Alcanivorax spp*. and *Acinetobacter spp*. This finding indicated that these hydrocarbon-degrading bacteria maintain substantial (not rare) populations *in-situ* based on growth on biogenic alkanes of algal or cyanobacterial origin, which are prevalent in the open, non-oil-contaminated ocean, and that the GoM may indeed be primed for crude oil biodegradation.

1.5 The Crude oil microbiome project

1.5.1 Documenting the diversity and ecology of oil degrading microbes on a global scale

The lack of a comprehensive database that integrates existing genomic/metagenomic data from oiled environments with physicochemical parameters known to regulate the fate of petroleum hydrocarbons currently limits data analysis and interpretations. While public repositories host massive amounts of data in the form of metagenomes or MAGs, currently no database exists to quantify or place these genomes in their environmental context with the *in-situ* physicochemical parameters. **Chapter 5** describes the curation of a comprehensive database documenting microbial indicators in the form of draft genomes

(MAGs, single-cell amplified genomes or isolate genomes) associated with petroleum hydrocarbons across the globe. All publicly available metagenome data associated with the DWH spill were binned into MAGs and included as a part of this database along with all publicly available MAGs associated with natural oil seeps, oil reservoirs, offshore oil platforms, previous oil spills etc. The biogeographical distribution patterns of these genomes as well as *in-situ* metadata of the samples that the genomes were recovered from, have been included as an interactive map as part of the webserver, which is called "Genome repository of oiled systems" or GROS (available at http://microbial-genomes.org/projects/GROS). The interactive webserver allows users to upload genomes (draft or complete) and the resulting data can help determine the taxonomic uniqueness of the queried genome, its relative *in-situ* abundance, the underlying physicochemical data of the samples that the genome is found to be present and the extent of association with oiled samples based on existing data. This project provides easy integration of data and capabilities to analyze new and unpublished data in the context of previously published data. The repository can be used to obtain a more holistic view of the microbial responses to oil, differences and similarities in the abundant/responding taxa across habitats, and identify biomarkers that can be universally representative (or not) of the different phases of oil biodegradation and ecosystem recovery. The curated database can also serve as a model approach for building similar resources for additional environmental processes and data of interest.

CHAPTER 2. A NOVEL BACTERIAL GENUS DOMINATES CRUDE-OIL CONTAMINATED COASTAL SEDIMENTS

Smruthi Karthikeyan, Luis M Rodriguez-R, Patrick Heritier-Robbins, Minjae Kim, Will A Overholt, John C. Gaby, Janet K. Hatt, Jim C. Spain, Ramon Rosselló-Móra, Markus Huettel, Joel E Kostka, Konstantinos T Konstantinidis

Originally published in ISME J on April 5th 2019, doi:10.1038/s41396-019-0400-5

2.1 Abstract

Modeling crude oil biodegradation in sediments remains a challenge due in part to the lack of appropriate model organisms. Here we report the metagenome-guided isolation of a novel organism that represents a phylogenetically narrow (>97% 16S rRNA gene identities) group of previously uncharacterized, crude-oil degraders. Analysis of available sequence data showed that these organisms are highly abundant in oiled sediments of coastal marine ecosystems across the world, often comprising ~30% of the total community, and virtually absent in pristine sediments or seawater. The isolate genome encodes functional nitrogen fixation and hydrocarbon degradation genes together with putative genes for biosurfactant production that apparently facilitate growth in the typically nitrogen-limited, oiled environment. Comparisons to available genomes revealed that this isolate represents a novel genus within the *Gammaproteobacteria*, for which we propose the provisional name "*Candidatus* Macondimonas diazotrophica" gen. nov., sp. nov. "*Ca.* M. diazotrophica" appears to play a key ecological role in the response to oil spills around

the globe and could be a promising model organism for studying ecophysiological responses to oil spills.

2.2 Background

The Deepwater Horizon (DWH) oil spill released over 780 million liters of oil and large amounts of natural gas (~1.7 x 1011 g) into the Gulf of Mexico and consequently, had a widespread impact on the pelagic, benthic and coastal ecosystems (Atlas and Hazen 2011, Bagby et al 2017, King et al 2015b). While most studies have focused on the fate of the oil in the deep sea plume and sediments (Handley et al 2017, Hazen et al 2010, Mason et al 2012), the impact of the DWH spill on coastal marine ecosystems remains comparatively understudied (Kostka et al. 2011). Following the spill, large amounts of weathered oil washed on shore contaminating an estimated 1773 km of the shoreline (Michel et al., 2013). Long-term effects of the DWH spill are still not well understood owing to the stochasticity and complexity of ecosystem processes as well as the lack of appropriate model microorganisms for studying the fate of oil in beach sands (King et al 2015a, Mendelssohn et al 2012, Michel et al 2013).

2.3 The ubiquitous uncharacterized *Gammaproteobacteria* in oiled sediments

Our previous studies revealed certain uncharacterized *Gammaproteobacteria* affiliated with the *Ectothiorhodospiraceae* family that showed high relative abundance in oil-contaminated sediments, exceeding the abundance of known hydrocarbon-degrading taxa (e.g., *Alcanivorax, Marinobacter*), especially during mid-to-late stages of degradation (Kostka et al., 2011; Rodriguez-R et al., 2015; Huettel et al., 2017). However, no cultivated members are available from this abundant *Gammaproteobacteria* group and thus, their

physiology remains unknown. In these studies, we also leveraged a metagenome time series to characterize the microbial community response to the DWH oil perturbation in beach sands (Pensacola Municipal Beach, FL). Our work revealed the succession patterns of individual microbial populations that responded to the spill up to one year after oiling when petroleum hydrocarbons were no longer detectable above baseline. Contrary to our expectations, we observed that generalist taxa, as opposed to specialists, were favored by the perturbation (Rodriguez-R et al 2015). Furthermore, PCR amplicon analysis of the nitrogen fixing genes (*nifH* gene) from these sands showed an increased abundance of nifH genes associated with various uncharacterized *Gammaproteobacteria* in the oil contaminated samples and returning to very low levels in the recovered sands (Gaby et al., unpublished). This was a potentially important finding since oil biodegradation is often nitrogen-limited, as exemplified by the addition of nitrogen fertilizer during cleanup efforts for the Exxon Valdez spill in Prince William Sound, Alaska (Pritchard et al 1992). A particular allele of *nifH* showed much higher abundance than the rest. In order to identify the full genomic context of this *nifH* gene and exact phylogenetic affiliation, targeted population reconstruction using visual inspection of the read coverage patterns of the assembly (See methods) yielded a draft metagenome-assembled genome (MAG-01) that included the abundant *nifH* gene allele. MAG-01's abundance increased from below detection levels in the clean/pre spill beach sand samples to ~30% of the entire microbial community in oiled samples, returning to low abundance levels in the recovered sediments (Figure 2-1, Appendix A, Figure A-1).

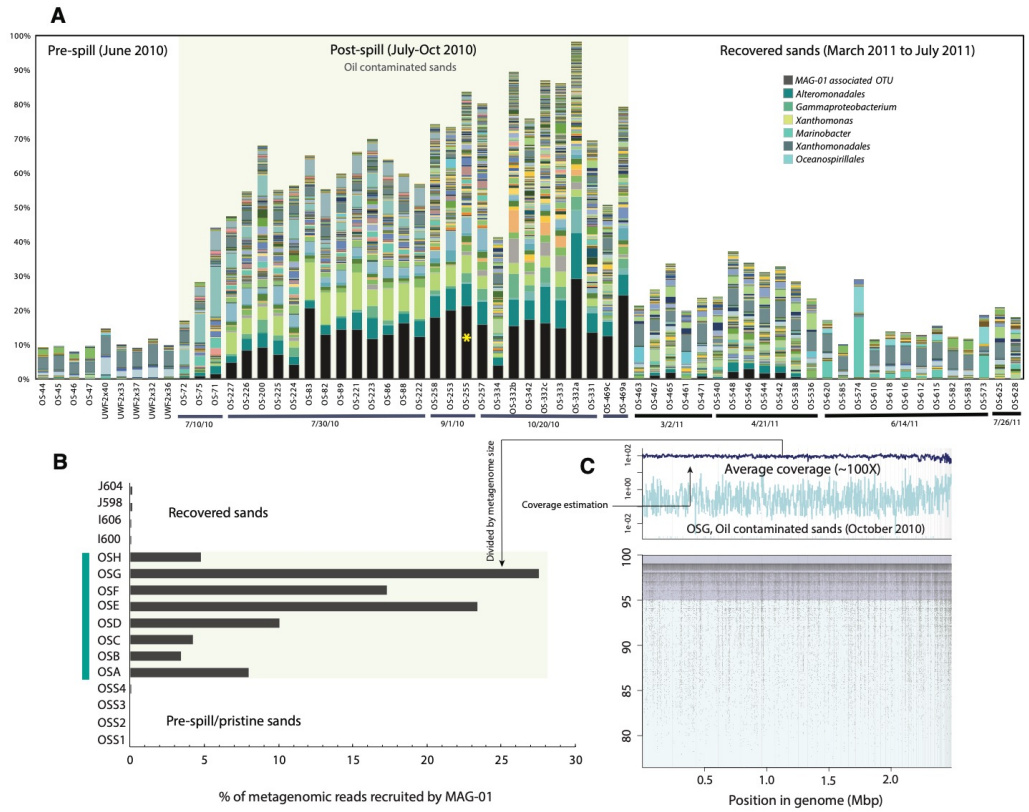


Figure 2-1: Relative abundance of MAG-01 in oiled and clean beach sands from Pensacola beach, Florida (USA). A: Abundance profiles of 16S rRNA gene-based OTUs detected in pre-oil, oiled and clean samples a year after the DWH oil spill. The MAG-01 16S-I is shown in black, at the bottom of the columns (denoted by an asterisk). Only the top 250 most abundant OTUs are shown. B: Average coverage, representing relative abundance, of MAG-01 sequence (x-axis) by the reads of the metagenomic datasets described in Rodriguez et al. 2015 (y-axis). C: (Bottom) Read recruitment plot showing where metagenomic reads of a contaminated sample (OS_G), which showed the highest abundance of MAG-01, mapped (x-axis) and their identity (y-axis). (Top) The dark blue histogram represents coverage, i.e., how many times each nucleotide base is covered by reads on average, by reads matching the reference MAG-01 sequence, at ≥80 bp in length

and ≥ 95% nucleotide identity, in 1000 bp-long windows; light blue represents reads matching at < 95% identity. The evenness of the coverage of the genome on the metagenomic datasets shows a sequence-discrete population. Note that the coverage values shown in panel B are derived from the average coverage obtained in the recruitment plots (dark blue histogram, Panel C) after normalizing for the size of the metagenomic dataset, and that the MAG-01 is not detectable in pre-spill samples and has low abundance in metagenomes of recovered microbial communities.

2.4 Targeted isolation efforts and metabolic versatility of the recovered isolate

Functional annotation of MAG-01 revealed putative genes for hydrocarbon degradation, nitrogen fixation, methanotrophy, urea metabolism, biosurfactant production, nutrient scavenging and other related processes that could enhance growth in oil-contaminated environments (Figure 2-2, Figure 2-3, Appendix A, Table A 3). Mass transfer limitations, nutrient (mainly nitrogen) and oxygen availability largely dictate the fate of the buried hydrocarbons and their bioavailability for microbial remediation (Head et al 2006, Huettel et al 2018, Kostka et al 2011). Hence, the functions identified were likely important for successfully coping with the oil perturbation and dominating the oiled microbial communities. Read recruitment plots of the metagenomes revealed an even coverage of MAG-01, at high nucleotide identity (>98%), indicative of a sequence-discrete population (Figure 2-1 and Appendix A, Table A 1) (Caro-Quintero and Konstantinidis 2012). However, its 16S rRNA gene was not assembled, as is common in binning efforts (Parks et al 2017), which prevented further taxonomic analysis. In an effort to identify the exact taxonomic affiliation of MAG-01 and further validate its genome sequence, visual

inspection of the MAG-01 assembly, complemented with PCR-walking for linking the rRNA operon, yielded a nearly complete genome.

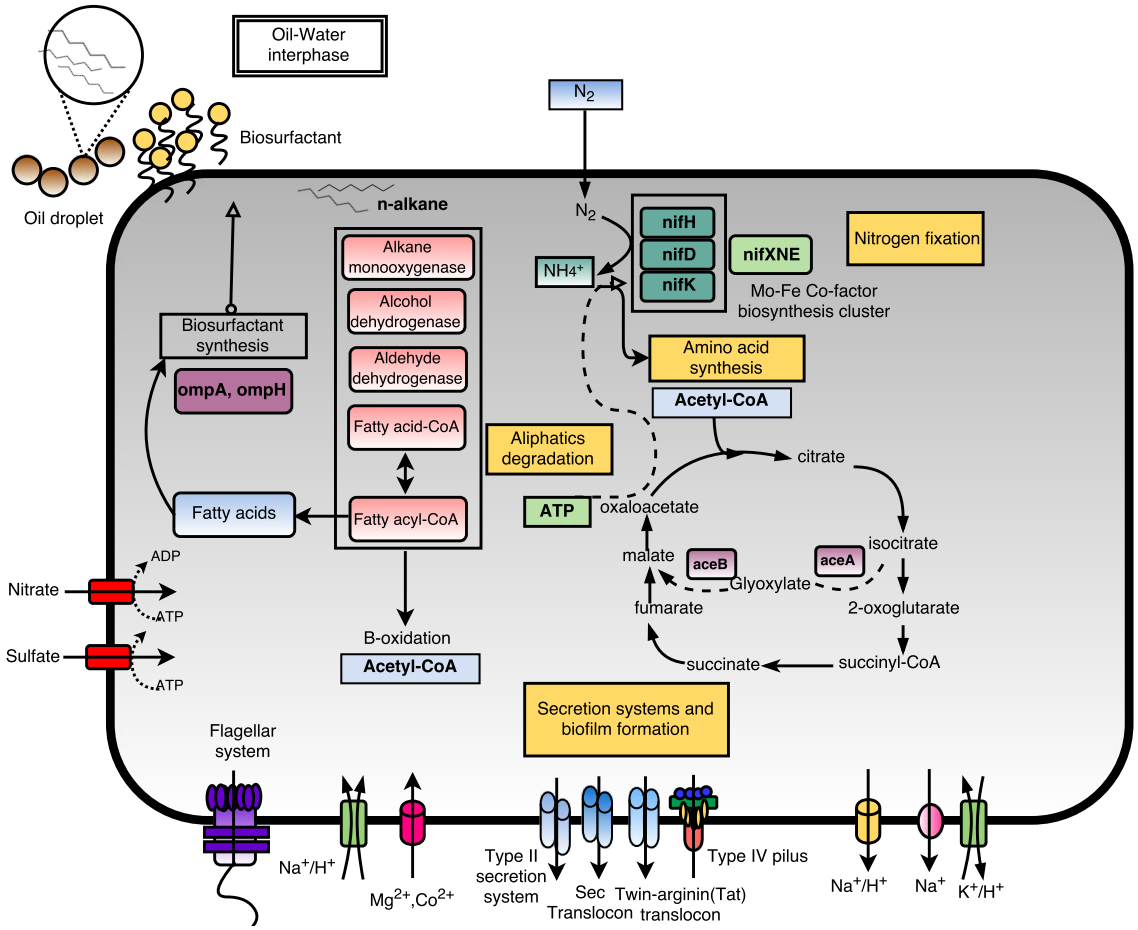


Figure 2-2: Schematic synopsis of the major metabolic and transport pathways encoded by KTK-01. Details on the hallmark genes (putative) identified for each pathway, including amino acid identity to their closest homolog, are shown in Appendix A, Table A 3.

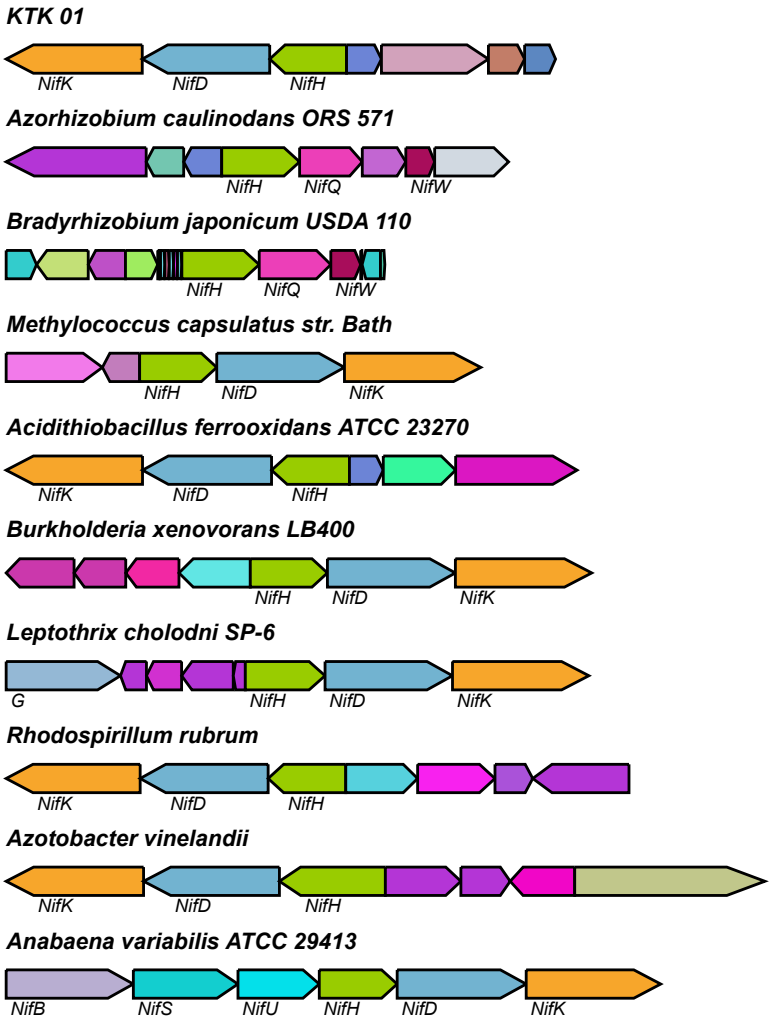


Figure 2-3: *nifHDK* operon of KTK-01 with its homology to other known *nifH* clusters. *Nif* operons encode for nitrogenase, the key enzyme responsible for conversion of dinitrogen gas to ammonia. Operon organizations were visualized using Gene Graphics (Harrison et al 2018).

To isolate the organism represented by the MAG-01 genome, enrichments were carried out using oiled sands collected at Pensacola beach as inocula in nitrogen-free minimal artificial seawater liquid media with no other added carbon source or nutrients.

Liquid aerobic enrichments were incubated for nine weeks before plating on agar plates containing the same (solidified) media supplemented with 0.2% (w/v) of source Macondo/MC252 oil as the sole carbon and energy source. Through colony PCR screening with the primers used in PCR walking above that were specific to MAG-01, the isolate KTK-01 was recovered, showing 100% nucleotide identity to MAG-01's 16S rRNA gene (see methods for further details). The genome sequence of KTK-01 showed 99.8% genome-aggregate average nucleotide identity (ANI) to MAG-01, revealing that it was a member of the natural population represented by MAG-01. Furthermore, several of the bioinformatically predicted functions mentioned above such as hexadecane degradation and nitrogen fixation were experimentally verified. KTK-01's *nifH* was also overexpressed in laboratory mesocosm experiments containing beach sands with added Macondo oil when compared to the un-oiled controls (Karthikeyan et al., unpublished data), indicating that it was functional under oiled conditions.

2.5 Ecological pervasiveness of the 16S rRNA gene sequence of the isolate

Screening of publicly available 16S rRNA gene amplicon or clone library datasets revealed a remarkable distribution of identical or almost identical (>97% nucleotide identity) sequences to KTK-01 in hydrocarbon-contaminated sediments of coastal ecosystems across the globe (Figure 2-4, Appendix A, Table A1). For instance, an operational taxonomic unit (I) identical to the 16S rRNA gene of KTK-01 (Figure 2-4) and metagenomic reads covering ~350X the KTK-01 genome (Appendix A, Figure A 1) was also of the most dominant OTUs/population in other beach sands affected by the DWH spill in the States of Louisiana and Alabama. Furthermore, organisms with KTK-01-like sequences were among the dominant taxa responding to other major coastal oil spills,

including the Prestige spill in the Galicia coast (>9% of total sequences) and in Cape Hallett in East Antarctica after an oil spill incident (Figure 2-4, Appendix A, Table A1). However, KTK-01-like sequences were below detection in the deep sea Macondo plume and oiled sediments, as well as in various uncontaminated water column metagenomes including those made available by the TARA Oceans expedition, underscoring an ecological niche specialization of KTK-01 in oiled, beach sands and coastal sediments.

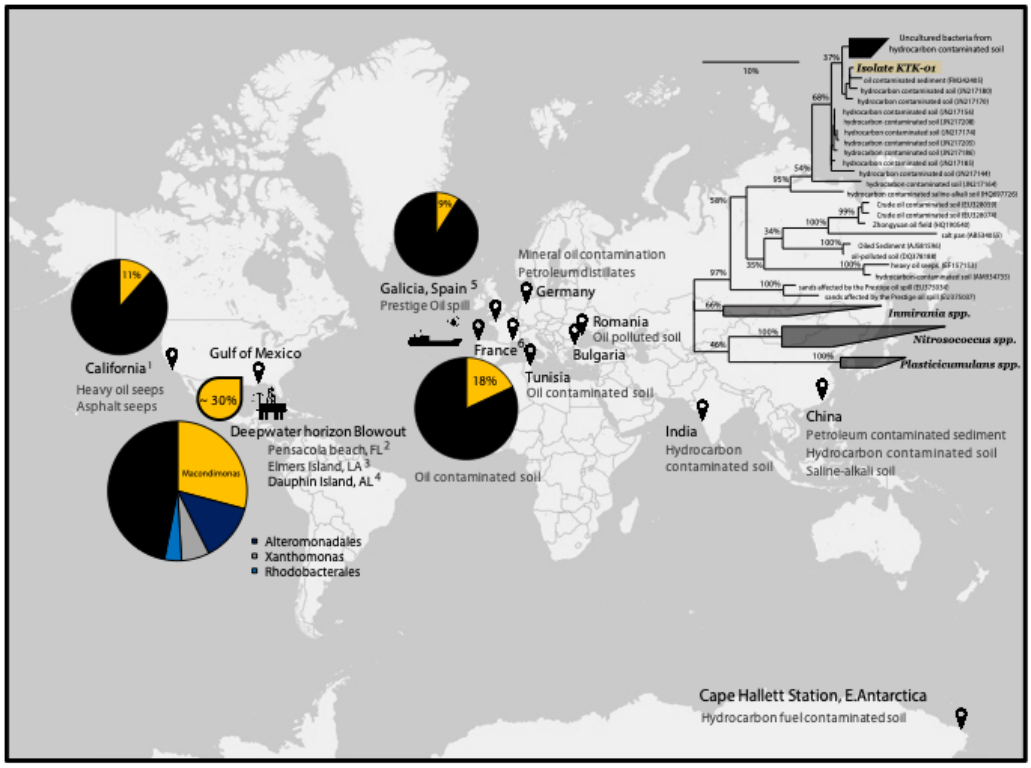


Figure 2-4: Phylogeny and distribution of KTK-01-like 16S rRNA gene sequences in oil-contaminated sites across the globe. Pie charts represent the fraction of total sequences showing > 97% nucleotide identity to the 16S rRNA gene sequence of KTK-01. For instance, 30% of the OTUs recovered from the beach sands impacted by Macondo oil

matched the 16S rRNA gene sequence at this level. Accession numbers of the datasets used are provided in Supplementary Table S1. Inset: 16S rRNA gene phylogeny of KTK-01 and selected close relatives. Maximum likelihood, as implemented in RaxML and using all homologous positions of the bacterial alignment in the LTP_123 dataset, was used to obtain the phylogenetic tree shown. Bootstrap values are indicated next to the branches. Complete 16S rRNA phylogeny is shown in Figure 2-7.

Simultaneous hydrocarbon degradation and nitrogen fixation by a single organism is rather uncommon among isolated hydrocarbon-degrading bacteria whose genomes were available to bioinformatically assess functional gene content (0/16; Appendix A, Table A 2). Furthermore, these genomes recruited almost no reads from available metagenomes of oiled coastal sediments, contrasting with the high abundance observed for KTK-01 (e.g., Figure 2-1). Therefore, our data indicated that the common practice of providing a nitrogen source during enrichment efforts might have biased the known diversity of cultivated hydrocarbon-degraders, and that nitrogen fixation is likely a strongly selected trait during oil biodegradation in-situ.

Collectively, our results indicated that KTK-01 represents a highly-promising model organism and a useful biomarker for the investigation of oil biodegradation in sediments, especially during mid-to-late phases of degradation (Figure 2-1). In addition, the phylogenetically distant affiliation of KTK-01 with its closest relative (classified) species, together with distinct genomic (e.g. AAI value < 65%, which corresponds to the genus level (Konstantinidis et al 2017)) and phenotypic traits (distinct diagnostic characters)

indicated that the new isolate could be classified as a new genus and species. The closest classified relatives based on the Microbial Genomes Atlas (MiGA) webserver (Rodriguez-R et al 2018) were autotrophic sulfur oxidizing species, *Thioalkalivibrio sulfidiphilus* (48.32% AAI) and *Thiohalobacter thiocyanaticus* FOKN1 (47.81% AAI) that do not fix nitrogen or degrade hydrocarbons like hexadecane. *Thioalkalivibrio sulfidiphilus* HL-EbGr7 is also the closest classified relative by 16S rRNA gene identity (91.8%) (Figure 2-5, Figure 2-6, Figure 2-7). For the new isolate we propose the name "*Candidatus* Macondimonas diazotrophica" gen. nov., sp. nov.

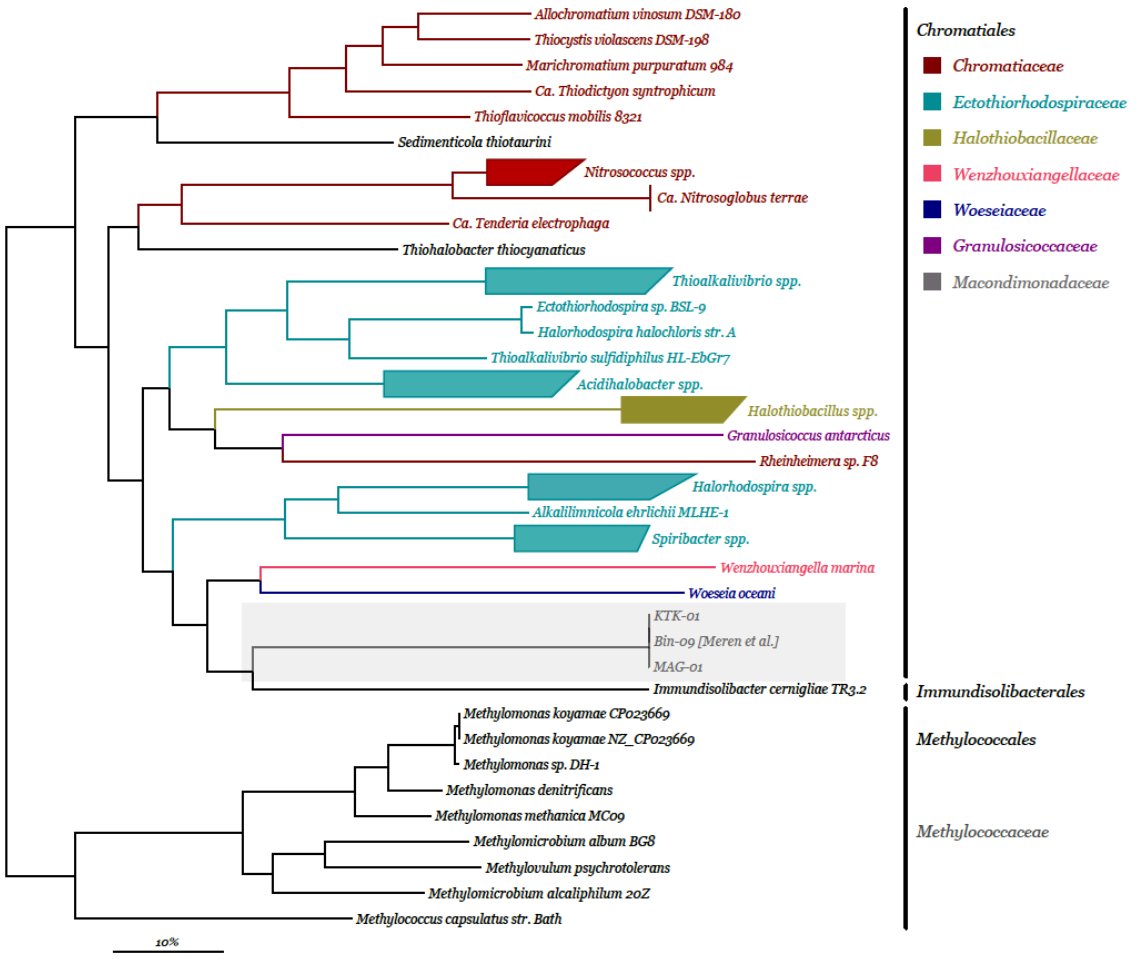


Figure 2-5: Universal gene-based phylogenetic reconstruction of genomes from four related orders of *Gammaproteobacteria*. 99 conserved single-copy gene sequences were used to reconstruct the phylogenetic relationships of available genomes from *Chromatiales, Immundisolibacterales,* and *Methylococcales* using FastTree with default parameters and the concatenated alignment of the 99 protein sequences. Members of the order *Chromatiales* are colored by family, and well resolved genera with multiple genomes available are collapsed for clarity. Collapsed clades are depicted as parallelograms with the length of the branch to the terminal nodes closer and farther from the root indicated by the

length of the bottom and top sides, respectively. Bin-09 as indicated in the Figure was obtained from (Eren et al. 2015).

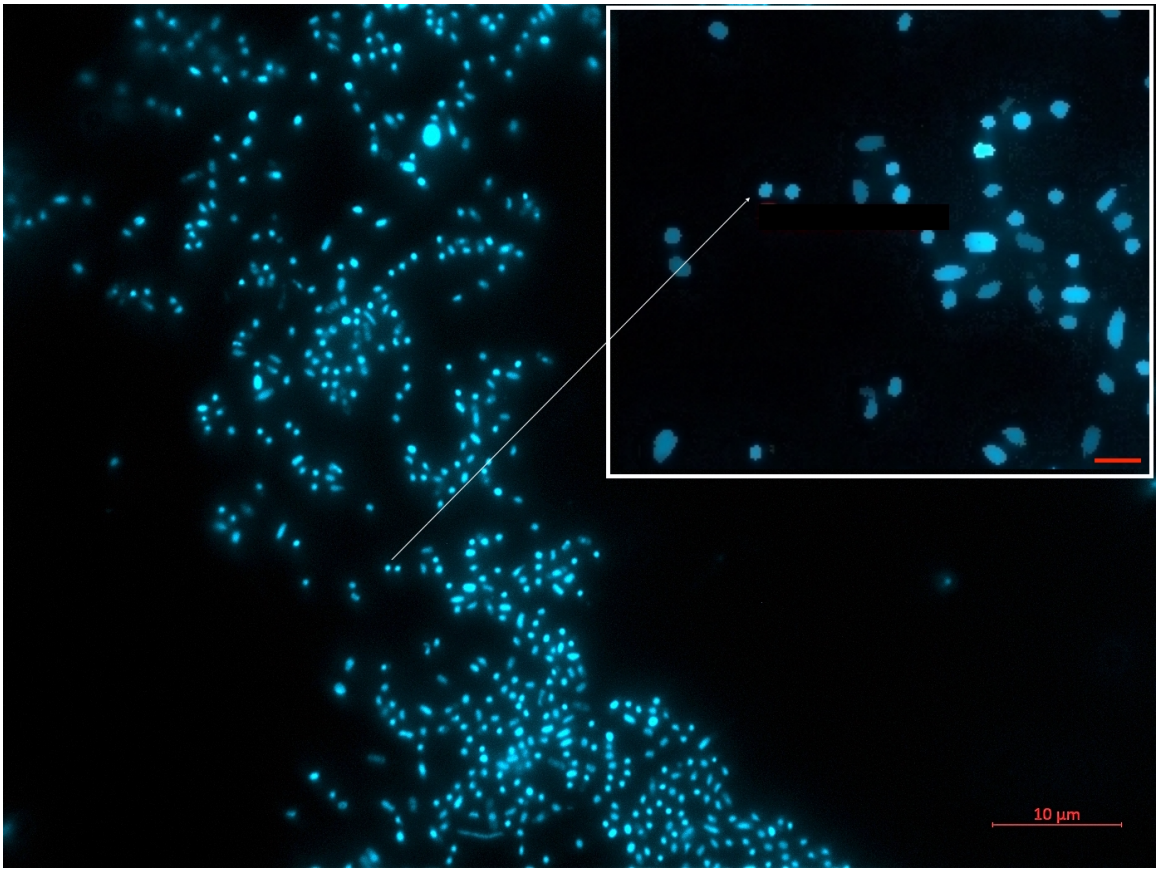


Figure 2-6: Fluorescence microscopy image of cells stained with DAPI. Inset scale bar 2 µM.

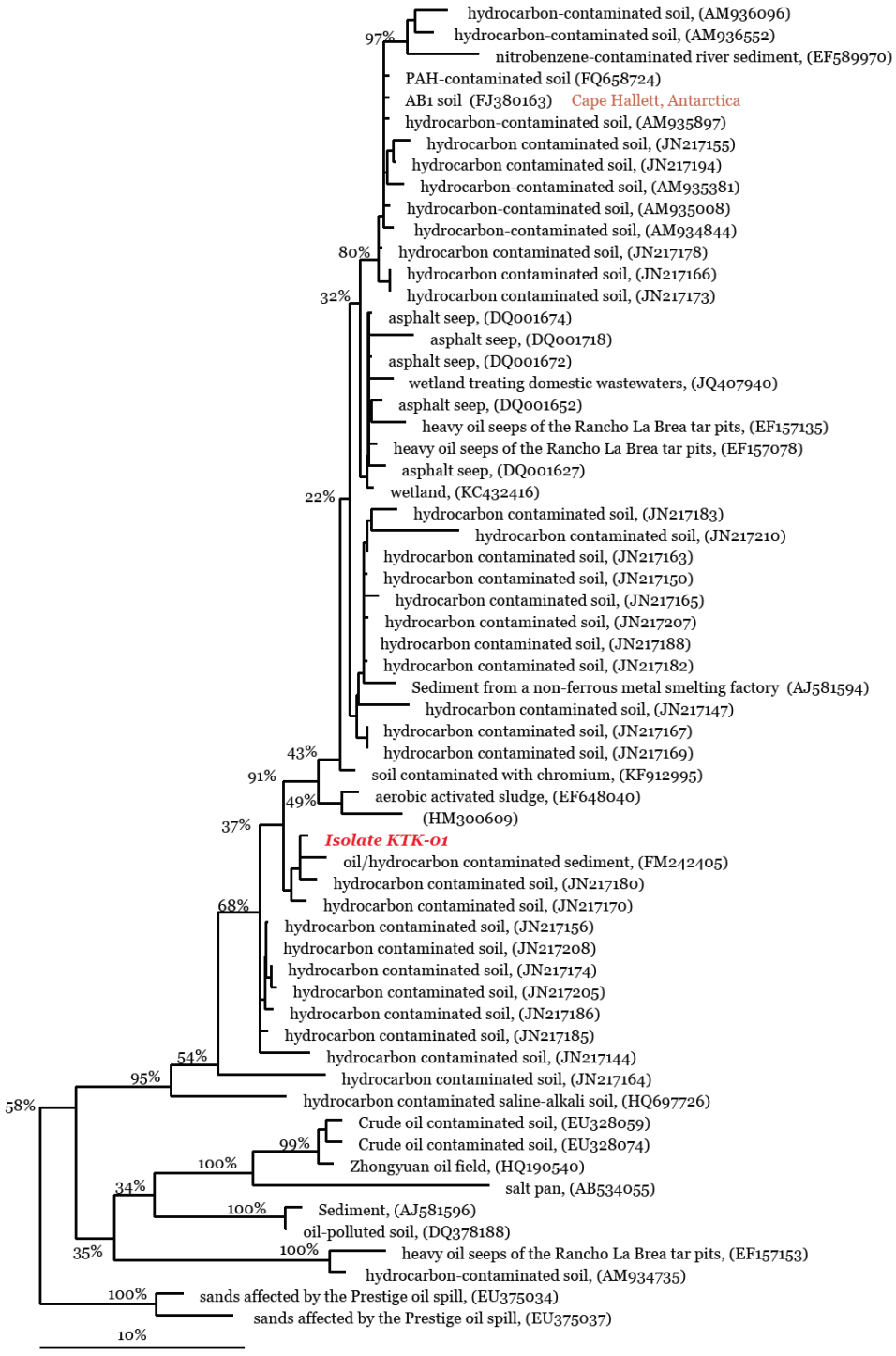


Figure 2-7: 16S rRNA gene phylogeny of KTK-01 isolate and close relatives. Phylogenetic tree reconstruction showing a consensus tree using different algorithms (neighbor joining and maximum likelihood) and different datasets and conservational

filters (implemented in the LTP_123 dataset) in where the most reliable topology is shown. The tree was constructed using the almost complete 16S rRNA gene sequence of the new isolate, that of the type strains of *Nitrosococcus oceani* (CP000127) and *Imnirania thermothiophila* (KT159732) present in the LTP_132 dataset, all their neighboring high quality sequences retrieved from the SILVA_REF_128 dataset, and the selected type strain sequences of the lineage closest to the new sequence's affiliation. The tree reconstruction is the result of the modification of a maximum likelihood RaxML tree reconstruction, using all homologous positions of the bacterial alignment and calculating their bootstrap values to show branching order stability. All bootstraps below 20% have been removed from the figure. Multifurcations indicate those main branches for which the order could not be solved. For the environmental sequences we have indicated their sample origin (given in their sequence entry) and the sequence accession number.

2.5.1 Description of "Candidatus Macondimonas" gen. nov.

Macondimonas, Ma.con.di.mo'nas. L. fem. N. monas, a unit, a monad; N.L. fem. N. Macondimonas, a monad from Macondo Prospect, the site of the Deepwater Horizon oil spill. Additionally, Macondo is a fictional town in *A Hundred Years of Solitude* by G. G. Márquez. In the book, the town of Macondo has a rapid population growth, a period of economic prosperity, and then a rapid population fall, which is reminiscent of the ecologic pattern observed for this group upon crude oil exposure. Members of this genus exhibit a coccobacilli morphology and a heterotrophic aerobic metabolism. No phototrophic, nor

chemoautotrophic growth or their corresponding genes in the genome were observed. The type species is “Ca. Macondimonas diazotrophica”.

2.5.2 Description of “Candidatus Macondimonas diazotrophica” sp. nov.

“*Ca. M. diazotrophica*”, di.a.zo.tro’phi.ca. Gr. Pref. di, in two; N.L. neut. N. azotum [from Fr. N. azote (from Gr. Prep. A, not; Gr. N. zôê, life; N.Gr. n. azôê, not sustaining life)], nitrogen; N.L. pref. diazo-, pertaining to dinitrogen; Gr. Adj. trophikos -ê -on, feeding, tending; N.L. fem. Adj. diazotrophica, one that feeds on dinitrogen, named after its ability to fix atmospheric nitrogen. Cells grown on solidified mineral artificial seawater media using hexadecane as substrate show a coccobacillus morphology, of about 0.6 μm in length and 0.35 μm in width, and formed circular colonies. Members of the species are aerobes, growing at a pH range of 6.5-8.5 with a pH optimum of 7.5, and a salinity range of 250-500 mM of NaCl, with an optimum concentration of 330 mM (or ~35 ppt). The range of temperature growth is 22-30ºC, with no growth observed at 4ºC and above 34ºC with an optimum 22-25ºC. Cells can grow with hexadecane and pyruvate as a sole carbon sources and fix nitrogen in the absence of an external nitrogen source. Genome size is ~2.8 Mbp with a G+C% content of 61.56.

2.6 Experimental Methods

2.6.1 Sample collection and DNA extraction

Beach sand samples were collected from the Pensacola Municipal beach (Florida, USA) before (pre-spill/clean) and after (oil contaminated and recovered beach sands) the oil slick reached the shoreline as described elsewhere (Rodriguez-R et al 2015). Sixteen

shotgun metagenomic datasets (Rodriguez-R et al 2015) and 122 16S rRNA gene amplicon datasets (Huettel et al 2018), sequenced from various sampling time points, were used in the analyses. Initial sample processing and sequencing were done as described previously (Rodriguez-R et al 2015)

2.6.2 Targeted reconstruction of the population genome

The effort to recover MAG-01 started with the identification of the organism(s) that encoded the most abundant *nifH* allele based on *nifH* gene amplicon data from oiled beach sand samples (Gaby et al., unpublished). To recover the population that carried the abundant nifH allele, the target allele was searched against the genes predicted in all assembled metagenomic contigs using Blast (Camacho et al 2009). The contigs containing genes identical to the target were extracted, which resulted in a collection of 8 contigs totaling 94 Kbp (training set). The contigs obtained were used to construct fragment recruitment plots from all the metagenomic datasets. The contigs had high and even sequencing coverage in the oil contaminated samples, ranging from 5 to 100X depending on the sample considered, and were virtually absent in the recovered and pre-spill datasets. This contig training set was used to identify other contigs with similar patterns of sequencing depth across datasets using the following Log-likelihood ratio statistic (D):

$$D(c) = 2(lnP'(c|X) - lnP'(c))$$

where *c* corresponds to the observed sequencing depth of the contig, and X corresponds to the set of sequences derived from the target organism. *P'(y)* is the Laplace smoothing ($\alpha=1$) of the empirical probability of the event *y*. *PI* is defined as the probability of observing the value of sequencing depth for contig *c* given the complete set of contigs, and the marginal

probability in *X* is defined as:

$$P(c|X) = 2 \times \min\left(ECDF(c|X), 1 - ECDF(c|X)\right)$$

Where *ECDF(c|X)* is the empirical cumulative distribution function evaluated at the sequencing depth of *c* derived from *X* (from the training set). All contigs from the assembly that resulted in log-likelihood ratio *D > -25* and similar tetra-nucleotide signature to the training contig set (Pearson's *r > 0.75*) were subsequently identified (about 6,000 contigs with a total length of 16 Mbp). The final assembly involved re-assembly of the contigs identified in the previous step in order to reduce redundancy (since they came from different metagenomes) using IDBA-UD (Peng et al. 2012), which resulted in a total of 122 contigs with a total length of 2.5 Mbp (N50 75 Kbp). The resulting draft population genome, MAG-01, had a CheckM (Parks et al. 2015) completeness of 96.39% and contamination of 0.32%. The likely taxonomic affiliation obtained for this bin using the MiGA webserver (Rodriguez-R et al. 2018) was a novel member of the *Gammaproteobacteria* order (novel genus with p-value 0.023, novel species with p-value 8e-4).

2.6.3 Recovering the 16S rRNA of the target MAG-01 genome

No 16S rRNA gene was recovered in the MAG-01 encoding the abundant *nifH* allele. To further study MAG-01 and assess its phylogenetic affiliation, we attempted to recover the 16S rRNA gene (16S) sequence. A 5S sequence was detected on one of the contigs of the MAG-01 using RNAmmer (Lagesen et al. 2007). To link the MAG-01 to its 16S, the 16S OTUs from metagenomes as well as the amplicon data (123 datasets) of the oiled and unoiled sediments from the beach sands were analyzed to identify which 16S I

correlated most strongly to the abundance of the MAG-01 based on the metagenome coverage (*i.e.*, strong correlation in relative abundances in the same samples was expected since all sequences originated from the same population). The metagenomic reads from each sample were mapped onto the MAG-01 in order to provide a coverage (relative abundance) estimate and compare to the relative I abundance from the 16S rRNA gene amplicon data from the same samples. The I with the highest correlation with to MAG-01 abundance (Pearson correlation coefficient > 0.9) was affiliated with an uncultured *Gammaproteobacteria* that likely belonged to the order *Chromatiales*.

The metagenome with the highest coverage was subsequently re-assembled using the IDBA-UD pipeline (Peng et al. 2012) (min kmer 25, max kmer size 121, step size 4) in an effort to yield a more complete genome. The 5S rRNA sequence identified in the MAG and ~400 bp of the 16S rRNA gene sequence was found on a contig in the assembled metagenome (100% nucleotide sequence identity). Another contig with similar abundances across metagenomes contained a partial 23S rRNA gene sequence (identified from the SSU and LSU database of SILVA). To retrieve the complete 16S rRNA sequence, a PCR walking experimental approach was implemented. A specific forward primer (5'-CGGATAAAGACGACCTGATT-3') was designed, ~390 bp upstream of the identified 16S sequence, and used in combination with a universal 16S rRNA reverse primer (1492R) to amplify the missing (not assembled) segment based on the DNA extracted from the oiled sediment sample with the highest relative abundance of MAG-01. Sanger sequencing of the amplified DNA segment provided a sequence that included about 350 bp of the upstream sequence (100% nucleotide sequence identity) as well as the available 400 bp of the 16S (100% identity), confirming our tentative scaffolding of the contigs described

above, and about an additional 1000 bp of the (missing) 16S rRNA gene sequence. The partial 16S rRNA gene sequence (1450 bp) was identified to represent MAG-01.

2.6.4 Estimating the abundance of the MAG-01 across the datasets

The MAG-01 genome (only contigs over 2 Kbp were used) was searched against all the reads from each sample using Blast. Fragment recruitment plots were constructed based on the Blast hits (threshold values: nucleotide identity ≥ 95% and alignment length ≥ 80 bp) using the enveomics collection (Rodriguez-R and Konstantinidis 2016b). The evenness of coverage of the reads at high nucleotide identity across the length of the reference genome sequence was used to evaluate the presence and discreteness of the population in each sample (Appendix A, Figure A 1). When coverage was even, the mean coverage across the genome was used as a proxy of *in-situ* relative abundance.

2.6.5 Bioinformatics functional prediction and data curation

Genes were predicted on the assembled contigs using MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010). The predicted genes were then searched against the SwissProt database using a Blastp (best match) search. The SwissProt IDs generated from the Blast results were mapped to their corresponding GO (Gene Ontology) terms and gene names using an in-house Perl script. Furthermore, RAST (Rapid Annotation using Subsystems Technology) was used for functional gene annotation (Aziz et al. 2008). To further validate these bioinformatics functional predictions, visual inspection of the alignment and data was performed for selected genes, especially against experimentally verified orthologs. The metabolic pathway reconstruction for the genome is shown in Figure 2-2 and the details of the identified hallmark genes of selected pathways are

provided in Appendix A Table A 1. Details of the *nifHDK* operon with its homology to other known *nifH* clusters are shown in Figure 2-3.

2.6.6 16S rRNA gene phylogeny

The complete 16S rRNA gene sequence of the strain KTK01 was compared to the sequences in the SILVA REF 132 database (Quast et al. 2013) and aligned against the universal alignment implemented in the ARB program package (Ludwig et al. 2004) using the SINA tool (Pruesse, Peplies, and Glockner 2012). Aligned sequences were first inserted into the preexisting tree of the SILVA REF 132 database to select the closest relative sequences. In addition, the new sequence was added to the LTP database (Yarza et al. 2010) that contains only curated 16S rRNA gene sequences of type strains of validly published species to select the closest relative type taxa. With the complete dataset of reference sequences, different trees were reconstructed using the neighbor joining and RaxML algorithms implemented in the ARB database. Different phylogenetic conservational filters (Munoz, Yarza, and Rosselló-Móra 2014) implemented in the LTP database as well as subsets of the reference sequences were used to find the optimal topology. The final tree shown in the manuscript is the result of a consensus of all tree-building approaches. Branching orders that could not be resolved were drawn as multifurcations (Figure 2-7).

2.6.7 Phylogenomics

All classified genomes available in the MiGA webserver, NCBI_Prok project (April/2018; n=11487 genomes in total), from the Proteobacteria orders *Chromatiales, Immundisolibacterales*, and *Methylococcales* were identified. From these genomes (n=48), we extracted all 101 single-copy, universal genes typically found in *Bacteria as* detected

by MiGA, selected genes present in at least 90% of the genomes (n=99), and aligned each gene independently using Clustal Omega with default parameters v1.2.1 (Sievers et al. 2011). Next, the concatenated alignment of the 99 genes was produced using Aln.cat.rb from the enveomics collection removing invariable sites (12,699 sites removed, 28,498 columns concatenated). Finally, phylogenetic reconstruction of the concatenated alignment was produced using FastTree with default parameters v2.1.7 (Price, Dehal, and Arkin 2010) (Figure 2-5).

2.6.8 Enrichment, isolation and growth characteristics

Enrichments were carried out using oiled sands collected from Pensacola beach (30°19'32.94"N, 87°10'29.68"W) on 20th October 2010. The sample, denoted by the metagenome ID OS_G, that represented the highest coverage of MAG-01 (making up ~30% of all assembled reads) was chosen as the initial inoculum for enrichments. The sample used as inoculum for the enrichments contained ~50 mg Total petroleum hydrocarbons (TPH)/kg of sand.

Based on the information from the analysis of the MAG-01 genome and functional potential, initial enrichment was carried out in two 150 ml Erlenmeyer flasks containing nitrogen-free minimal artificial seawater medium (with 1ml/L each of vitamin mixture, B12 solution and Thiamine solution) (Widdel and Bak 1992). Each flask contained 5g oil contaminated sand (inoculum) in 50 ml of medium. These incubations and all subsequent growth experiments were carried out aerobically shaken at 150 rpm, at room temperature, and in the dark. After about 9 weeks of incubation, serial dilutions of the above enrichment cultures were spread onto plates containing solidified minimal N-free artificial seawater

medium supplemented with 0.2% (w/v) of source oil (Macondo oil) as the sole carbon source. The oil used was previously sterilized as described by Widdel and Bak (Widdel 2010). Noble agar was used as the solidifying agent to eliminate trace nitrogen and other impurities. Five distinct colonies (based on their morphology/visual appearance) were observed on the plate and were chosen for further analysis. Colony PCR for the 16S rRNA gene was performed using universal forward and reverse primer sets (8F, 5'-AGAGTTTGATCCTGGCTCAG-3' and 1492R) on the 5 distinct colonies. To identify the (target) isolate, the same specific forward primer used for PCR walking that was described above was used in colony PCR along with the universal 1492R reverse primer and (successful) reactions were screened by agarose gel electrophoresis. Isolates producing amplicons were chosen for Sanger sequencing (Genewiz sequencing facility, South Plainfield, NJ) and further analysis. 16S rRNA of one of the 5 isolates obtained matched 100% with that recovered from MAG-01. The 16S rRNA of the other 4 screened colonies were close relatives (>97% nucleotide identity) of *Kangiella geojedonensis* NZ CP010975, *Marinobacter hydrocarbonoclasticus* ATCC 49840, *Microbulbifer agarilyticus* NZ CP019650 and *Loktanella hongkongensis* strain KNUSC1020. Further tests and purification of isolate KTK-01 colonies were carried out in N-free and sediment free minimal seawater media using hexadecane as the sole carbon source (0.2-2% v/v for liquid culture). Hexadecane was chosen as preferred carbon source since the bioinformatics functional annotation of MAG-01 revealed genes for degradation of medium to long chain alkanes. Furthermore, based on our previous study of the metagenomes from Pensacola beach, MAG-01 appeared to more enriched at the later stages of oil degradation which had relatively lower concentrations of the lighter alkanes (i.e., compounds lighter than C12).

The colony morphology of isolate KTK-01 grown in minimal marine salts media supplemented with 0.2% hexadecane as the sole carbon source is shown in Fig S6 revealing a coccobacilli morphology. The other 4 colonies did not grow in the N-free minimal seawater media supplemented with hexadecane. The optimal growth conditions of the KTK-01 isolate were evaluated over a range of salinities (170, 250, 330, 500, 660 mM of NaCl), pH (5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0 and 8.5) and temperatures (4-40 ˚C). For this, the cultures were grown in triplicate in modified media as described above with pyruvate (0.5 g/L) as the carbon source and ammonium chloride (1 g/L) as the N source (Kostka et al. 2011) in order to facilitate fast growth. Growth was quantified by measuring the optical density (600 nm) of the cultures at periodic intervals. The optimal growth conditions of the KTK-01 isolate were evaluated over a range of salinities (170, 250, 330, 500, 660 mM of NaCl), pH (5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0 and 8.5) and temperatures (4-40 ˚C). For this, the cultures were grown in triplicate in modified media as described elsewhere with pyruvate (0.5 g/L) as the carbon source and ammonium chloride (1 g/L) as the N source (Kostka et al. 2011). Growth was quantified by measuring the optical density (600 nm) of the cultures at periodic intervals.

2.6.9 Acetylene reduction assay for nitrogen fixation potential

Nitrogen-fixation rate potential of the KTK-01 isolate was determined using an acetylene reduction assay (Capone 1993). Cultures (5 ml) were grown in Jensen's nitrogen free media (HiMedia Laboratories, India) in 30 ml sealed Balch tubes. Controls included *Azotobacter vinelandii* (positive), *Escherichia coli* (negative) and autoclaved samples (killed controls). Five percent of the headspace in the tubes was replaced with acetylene and all the samples were incubated in the dark on a shaker at 180 rpm. Growth was

monitored periodically by measuring the optical density (600 nm). A Gas Chromatography – Flame Ionization Detector (Greenhouse Gas Monitoring GC, SRI Instruments, Torrance, California, USA) equipped with dual 2 m HayeSep-D columns was used to quantify ethylene production. Samples were measured routinely on the GC for ethylene production. All incubations were carried out in triplicate. Ethylene peaks were observed only in the positive controls and the isolate of interest. No ethylene peaks were observed in the negative and killed controls.

2.6.10 Whole genome shotgun sequencing

DNA was extracted from an active culture of KTK-01 growing on 2% hexadecane during late exponential phase (~0.9 of OD 600 nm) using the QIAmp DNA Mini kit (Qiagen, MD) according to the manufacturer's protocol. DNA sequencing 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. Library concentrations were determined by fluorescent quantification using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific) and samples run on a High Sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine library insert sizes. The KTK-01 isolate genome was sequenced using a MiSeq reagent v2 kit for 500 cycles (2 x 250 bp paired end run) on an Illumina MiSeq instrument (located in the School of Biological Sciences, Georgia Institute of Technology). Adapter trimming and demultiplexing of the sequenced samples were carried out by the MiSeq Control Software.

2.7 Data Availability

The data reported in this paper is publicly available through the Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC), under the accession numbers R5.x278.000:0014 and R5.x278.000:0002. The genome sequences have been deposited in NCBI and are available under BioProject PRJNA530182. The metagenome sequences as well as the assembled genome sequences are also available at http://enve-omics.ce.gatech.edu/data/.

2.8 Acknowledgements

The authors would like to thank Dr. Aharon Oren for his valuable input on the naming of the new type species, and the Gulf of Mexico Research Initiative (GoMRI) for supporting this work through Grant No. 231611-00 (RFP V). RRM acknowledges the financial support for his sabbatical stay at Georgia Tech from the Spanish Ministry of Sciences, Innovation and Universities (Grant No. PRX18/00048).

CHAPTER 3. INTEGRATED OMICS APPROACHES UNRAVEL COMPLEX RESPONSE OF INTERTIDAL SEDIMENTARY MICROBIAL COMMUNITIES TO THE DEEPWATER HORIZON OIL SPILL

Smruthi Karthikeyan, Minjae Kim, Patrick Heritier-Robbins, Janet K. Hatt, Jim C. Spain, Will A. Overholt, Markus Huettel, Joel E. Kostka and Konstantinos T. Konstantinidis

3.1 Abstract

Crude oil buried in intertidal sands may be exposed to alternating oxic and anoxic conditions. The effect of these tidally-induced biogeochemical oscillations is not understood, impeding remediation and managing efforts after oil spills. Here, we used a combination of metatranscriptome and genome-resolved metagenomics to elucidate the controls of microbial oil degradation undergoing temporal, aerobic-anoxic cycles simulated in laboratory chambers. Oil biodegradation rates were 5-fold higher during the oxic phase compared to the anoxic one, with a relatively constant ratio between aerobic and anoxic oil decomposition rates, even after prolonged anoxic conditions. Microbial activities during the oxic phases promoted oil biodegradation by re-oxidizing available electron acceptors and by providing nitrogen, an oil-limiting nutrient, through biological nitrogen fixation. Based on metatranscriptomics analysis, genes for sulfide oxidation were expressed also during anoxic phases, indicating use of alternative electron acceptors for

sulfide detoxification. We found that all population genomes reconstructed from the laboratory experiment were also present, typically in low abundances (rare biosphere), in metagenomic data from field samples in Florida, including the anoxic-enriched genomes, indicating that the intertidal communities are "adapted" to changes in redox conditions. These results help explain the relatively fast biodegradation of DWH oil in sandy sediments observed previously, and have implications for managing future oil spills.

3.2 Introduction

The Deepwater Horizon (DWH) oil spill was the largest accidental marine oil spill in history, affecting benthic ecosystems over vast areas of the Gulf of Mexico (King et al. 2015). The complexity of modelling the deep-sea plume dynamics, coupled with the large amounts of natural gas that were released, made it challenging to track the fate of the oil (Camilli et al. 2010; Murray, Boehm, and Prince 2020). It has been estimated that approximately 50% of the oil released from the DWH blowout reached the ocean surface and a significant portion of this was transported to coastal ecosystems and buried in the sediments (King et al. 2015; Rodriguez-R et al. 2015). The *in-situ* conditions namely, temperature, oxygen availability, pressure, degree of weathering of oil, were vastly different across the ecosystems impacted by the spill thereby creating diverse ecological niches harboring specific microbial signatures (Kimes et al. 2014). Biodegradation mediated by a complex network of microorganisms ultimately controls the fate of the majority of oil hydrocarbons that enter these ecosystems (Atlas and Hazen 2011; Hazen et al. 2010; Head, Jones, and Röling 2006; Leahy and Colwell 1990).

However, microbial biodegradation of oil under anoxic conditions prevailing in submerged nearshore sediments and beaches can be extremely slow (Shin et al. 2019), resulting in the accumulation of petroleum hydrocarbons. The burial of oil can restrict access to O_2 such that the oiled layers become anoxic, leading to almost imperceptible rates of biodegradation (Prince 2010). Oil can persist in largely un-weathered forms in such anoxic sediments and thus, may contain relatively large concentrations of harmful polycyclic aromatic hydrocarbons (PAHs) that can be released to the surface during storm events. Furthermore, sediment-oil-agglomerates (SOA) and surface residue balls (SRBs), which are oil-sand aggregates, also washed up along the shorelines of the affected areas and were buried under coastal sediments. It has been estimated that the complete biodegradation of these buried SOAs can take decades (Bociu et al. 2019). Similarly long times for complete oil biodegradation were observed after the Exxon Valdez spill (Alaska, USA). After this spill, large amounts of oil were washed up and buried in shoreline sediments where they persist despite extensive biostimulation efforts (Short et al. 2004). Much less information is available on hydrocarbon degradation under anoxic conditions and an improved understanding would enable better predictions on the fate and transport of oil.

Although certain microbial community successional patterns have been established in response to hydrocarbon availability (Rodriguez-R et al. 2015; Mason et al. 2014; Kimes et al. 2014; Kostka et al. 2020), there is still a limited understanding of the universal applicability of these patterns, and how these patterns may differ depending on the specific environmental conditions and water transport processes. For instance, the effect of oxygen oscillations that occur regularly as an effect of tidal current patterns in the water saturated

(subtidal) and temporally saturated (intertidal) sediments on the microbially-mediated oil degradation is unknown. Studies modeling the degradation of petroleum hydrocarbons in the capillary fringe showed transverse dispersion and diffusion essentially controlling biodegradation (Luo et al. 2015). High organic decomposition rates in sandy intertidal zones have been attributed primarily to rapid advective transport (rather than diffusion) of oxygen into the sediment (Jorgensen 1994; Forster, Huettel, and Ziebis 1996). Tidal pumping can enhance gas transport rates in dry upper intertidal and supratidal zones (Huettel et al. 2018). As our experiments addressed microbial oil degradation in sandy sediment, we focused on advective transport of oxygenated water in our incubations. Our previous study elucidating microbial community succession patterns from samples recovered from Pensacola Beach (PB), FL, USA after the DWH oil washed onshore revealed specific microbial community signatures corresponding to the stage of oil degradation (Rodriguez-R et al. 2015). Redox oscillations have been previously implicated in controlling the degradation of natural organic matter (Aller 1994; Wakeham and Canuel 2006) where the typical oxic-anoxic cycles in bioturbated sediments lead to more rapid organic matter decomposition than under constant conditions or unidirectional redox change. Residence times were typically longer in the anoxic (~10x) phases compared to the oxic phases. However, the effect of alternating periods of aerobic and anoxic conditions on buried hydrocarbons and the associated microbial community dynamics have not been studied. As tidal oscillations can typically operate across a range of time scales, this study focused on neap-tide cycles which typically average 14 days in order to establish effects of prolonged aerobic-anoxic conditions on the microbial communities as well as assess the potential accumulation of metabolites (sulfides) from prolonged anoxic conditions. At low

concentrations observed in tidal sands, sulfides present a potent electron donor or substrate for microbial growth. Thus, sulfides can be either a substrate for growth or a potential toxin. Furthermore, the cycles were designed to assess the timeline for depletion of oxygen through microbial activity (i.e. time taken to establish anoxic conditions in these oiled sediments).

Closed-system studies under controlled conditions in the laboratory can help untangle some of the complex microbial interactions occurring in permanently or periodically anoxic sediments and thus, provide new insights into the effects of redox oscillations in these sediments. However, laboratory-based studies generally do not replicate well the *in-situ* gradients and accordingly, their results do not match field results. In the present study, we used laboratory advection chambers that closely simulate the sandy beach intertidal zone, including oxygen and water transport phenomena. Oxic-anoxic cycles were physically induced in these chambers and the activities of the responding microbial communities were monitored through a combination of time-series metagenomics and metatranscriptomics along with hydrocarbon fingerprinting techniques. Metagenome-assembled genome (MAG) data revealed specific populations linked to specific stages of oil biodegradation mimicking the same patterns observed in the field, with high degree of similarity of the MAGs obtained. Metabolic pathway reconstruction along with metatranscriptomic and hydrocarbon fingerprinting analysis revealed a highly dynamic microbial community adapted to the switching oxygen regimes. Metabolic pathway analysis, combined with hydrocarbon compositional analysis and species abundance data, revealed substrate specialization that explained the successional pattern of oil-degrading bacteria. These results advance our understanding of the fate and

consequences of deposited DWH crude oil in benthic ecosystems and the *in-situ* microbial community response.

3.3 Experimental Methods

3.3.1 Experimental set-up and operation

Beach sand was collected from the intertidal zone in Pensacola Beach, PB (30°119.57 N, 87°110.47 W) and incubated in six mesocosm chambers as described elsewhere (Ehrenhauss and Huettel 2004; Huettel and Rusch 2000). Three of the chambers were amended with 5mg/g of weathered Macondo oil and served as biological triplicates while three chambers served as controls (no oil added). Each chamber was equipped with an online dissolved oxygen sensor (PreSens GmBH, Regensburg, Germany) for continuous monitoring of oxygen and temperature. The mesocosms were operated as closed systems and in the dark throughout the incubation. The six chambers were also equipped with air pumps to control the dissolved oxygen levels. All six chambers were aerated to full saturation at the start of the experiment after which the air pump was turned off. The first sample (corresponding to the 1st anoxic phase) was taken 18 days after the onset of the incubation, when oxygen levels were dropped to essentially 0% saturation. A glass sediment corer was used to sample the sands from multiple of the chamber in order to minimize errors due to sample heterogeneity (i.e., differential distribution of oil in the sediment/chamber). The samples were flash frozen in liquid nitrogen and stored at -80 ^{0}C for further nucleic acid extraction and total petroleum hydrocarbons (TPH) analyses. The pore water was sampled for nutrient concentrations (ammonia, nitrite, nitrate, sulfate) and analyzed using the Hach spectrophotometer (Hach Instruments, Colorado, USA). TPH,

hydrocarbon fingerprinting analyses and moisture analyses were carried out at Eurofins Lancaster Laboratories (Lancaster, PA, USA). After the 1st sampling point, the system was re-aerated to 100% air saturation and the oxygen depletion rate was subsequently monitored. The 2nd sampling point was taken when the oxygen levels dropped to 50% saturation corresponding to the aerobic phase sample. This cycle was repeated to generate four sampling timepoints (two oxic and two anoxic cycles), excluding the time zero (baseline) sampling time (Figure 3-1 A and B).

3.3.2 Community DNA and RNA extraction

Community DNA was extracted from the sand samples using the Power Soil kit (Qiagen Inc., MD) using the manufacturer's protocol. Community RNA was extracted using a modified protocol of the Dneasy PowerMax Soil Kit (Qiagen Inc., MD). Details on the RNA protocol have been provided in the Supplemental documentation. The extracted RNA was treated with Turbo Dnase, 2U/L (Invitrogen) and the RNA integrity verified on an Agilent Bioanalyzer 2100 instrument using the RNA Pico6000 chips. cDNA libraries were prepared using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Illumina, San Diego, CA, USA), generating libraries with average insert sizes of 381 bp. DNA and RNA concentrations were quantified using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific). 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. 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 HiSeq instrument.

3.3.3 Bioinformatic analysis of metagenomic and metatranscriptomic data sets.

The sequenced shotgun metagenome and metatranscriptome reads were trimmed and quality checked using Solexa QA(Cox, Peterson, and Biggs 2010). Only reads having a Phred score of 20 (>99% accuracy per base-position) and trimmed read lengths of 50bp or longer were retained for downstream analysis.

3.3.3.1 Community diversity estimates

Sequencing coverage of the metagenomic datasets was estimated using Nonpareil (Rodriguez and Konstantinidis 2014). Whole-community, pairwise distances were calculated using MASH, a tool employing the MinHash dimensionality reduction technique to compare sample-to-sample sequence composition based on k-mers (Ondov et al. 2016). The MASH-based distance matrix was visualized using principal coordinate analysis (PcoA) and non-metric multidimensional scaling (NMDS) plots.

3.3.3.2 Recovery of 16S rRNA gene fragments from shotgun metagenomes

16S ribosomal rRNA (16S) gene fragments were extracted using Parallel-META 2.0 (Su et al. 2014) and taxonomically classified using the RDP database (Cole et al. 2014). Recovered 16S fragments were clustered ('closed-reference I picking' strategy) using UCLUST (Edgar 2010) and taxonomically classified using best match hits in the RDP database at an ID ≥ 97% in QIIME (Kuczynski et al. 2011).

3.3.3.3 Functional annotation of short reads

FragGeneScan (Rho, Tang, and Ye 2010) was used to predict protein coding regions in short-reds using the 1% Illumina error model. The predicted protein-coding regions were searched against the curated Swiss-Prot (Bairoch and Apweiler 2000) protein database using Blastp (Camacho et al. 2009). Only matches with a bitscore >60 and amino acid identity >= 40% were used. The Swiss-Prot database identifiers were mapped to their corresponding metabolic function based on the hierarchical classification subsystems (SEED subsystems) (Overbeek et al. 2014). For estimating the abundance of the nitrogen cycling genes, ROCker models (Orellana, Rodriguez-R, and Konstantinidis 2016) were employed, and the number of ROCker identified reads were transformed to genome equivalents (i.e., how many of the total cells encode the gene of interest) using the abundance of reads encoding the single copy *rpoB* gene for comparison as described elsewhere. Alkane monooxygenase (*alkB*) gene diversity was estimated based on the short reads as described previously (Rodriguez-R et al. 2015).

3.3.3.4 Assembly and population genome binning

Co-assembly of the short reads form the biological triplicates was performed using IDBA-UD (Peng et al. 2012) and only resulting contigs longer than 500 bp were used in further analysis (functional annotation, binning and MyTaxa classification). MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010) 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(Tsementzi et al. 2019) . The bins were checked for quality (contamination and

completeness) using CheckM (Parks et al. 2015) 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 performed using fragment recruitment plots that were generated with the scripts of the enveomics collection (Rodriguez-R and Konstantinidis 2016a) as previously described (Konstantinidis and DeLong 2008).

3.3.3.5 Functional annotation of MAGs

Protein-coding regions of MAGs were searched against the Swiss-Prot (Bairoch and Apweiler 2000) 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 genes mapping to each function was subsequently estimated based on the number of predicted genes from each MAG assigned to the function. Selected functions of interest such as biomarker genes of oil degradation were manually verified by examining the Blast output and/or the conservation of genes in the same operon.

3.3.3.6 Metatranscriptome analysis

The trimmed transcriptomic reads were processed using SortMeRNA v2.0 (Kopylova, Noé, and Touzet 2012) to identify and remove ribosomal RNA reads (no rRNA subtraction before library creation and sequencing was attempted due to relative low RNA yields from control samples). The protein coding regions were predicted on the non-ribosomal-encoding reads using Prodigal 2.0 (Hyatt et al. 2010) and these reads were annotated using the SwissProt database as described above. The differential expression of genes was evaluated using the DESeq2 (Love, Huber, and Anders 2014) package in R as

described by Kim et al., 2018 (Kim et al. 2018). A log 2-fold threshold was used to evaluate the genes significantly expressed in one treatment vs. another (corresponding to an adjusted p-adj value of 0.01 or lower).

3.4 Results

3.4.1 Microbial community successional patterns mirrored those of field samples

The advective flow chambers were operated with temporal oxic-anoxic cycles for a period of 137 days. Initially, all six chambers were saturated with air via air pumps. Samples were taken at the start of the experiment (oxygen saturation conditions) and at appropriate intervals thereafter. The typical duration of the oxic cycle was about 12 days, followed by anoxic conditions for similar duration, four cycles were performed (Figure 3-1). The air pump was turned off after air saturation and dissolved oxygen (DO) consumption was measured over time. Increases in DO consumption rates increased with successive cycles. Over the course of the incubation, 88% reduction in the total petroleum hydrocarbons (TPH) and 68% reduction in polycyclic aromatic hydrocarbons (PAHs) were observed (Figure 3-1). Oil fingerprinting analysis revealed an approximately 5-fold faster oil degradation rate under oxic vs. anoxic conditions, over the course of the incubation. In total, 21 metagenomic data sets, including biological replicates for each sampling condition and time point, and 15 metatranscriptomic datasets totaling ~200 Gbp of sequencing data were generated (Appendix B, Table B 1) to study the microbial responses and gene activities during the incubation. An estimated abundance-weighted average coverage of the oiled sediments using Nonpareil (Rodriguez and Konstantinidis 2014) ranged from 60-75% and that of control samples ranged between 45-70% (Figure 3-1 E, Appendix B, Figure B

1), indicating that our sequencing effort was not limiting for assembly and population genome binning.

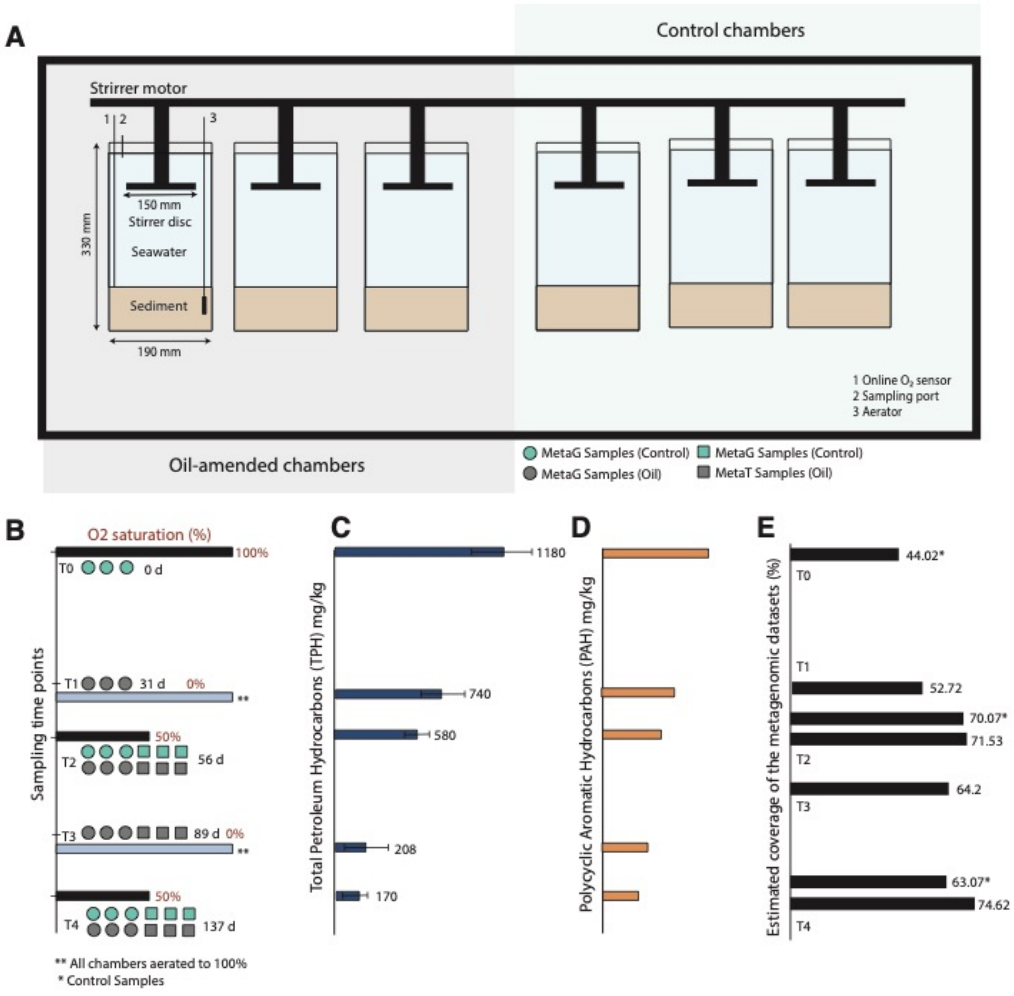


Figure 3-1: Mesocosm experimental design and operation. A. Graphical representation of the oil-treated and control advective flow chambers. Three chambers (replicates), marked as "Oil-amended", were incubated with 5 mg/g of weathered Macondo oil and another three served as controls. B. Sampling timepoints and type of analysis performed. Circles represent metagenomic samples taken and squares represent metatranscriptomes.

C and D: Abundance of the total petroleum hydrocarbons (TPH) and polycyclic aromatic hydrocarbons (PAH). E. Sequencing coverage estimation of the metagenomic datasets using Nonpareil. The control samples are indicated with an asterisk (*).

Community composition analysis based on 16S rRNA gene fragments recovered in the metagenomes revealed that the control chambers were remarkably stable throughout the incubation. In contrast, the oiled chambers underwent a drastic microbial community succession from an initial dominance of gammaproteobacterial lineages to an alphaproteobacterial-dominated community over time. Notably, the anoxic phases had an increased abundance of *Deltaproteobacteria* (predominantly sulfate reducing bacteria; Figure 3-2A). The non-metric multidimensional scaling (NMDS) plot constructed based on the MASH distances among whole metagenomes revealed clear clustering patterns based on treatment (Figure 3-2B). Specifically, the clustering of control samples indicated a relatively stable community, whereas the oiled sediment samples clustered separately according to incubation time and treatment (i.e., aerobic vs. anoxic). Moreover, the metagenome samples of (treated) oxic and control metagenomes clustered with the oiled and pre-oil (clean) sand metagenomes, respectively, when compared to previous field investigations of the municipal beach in Pensacola (Florida, USA), where sands for the mesocosms were collected (Rodriguez-R et al. 2015). Furthermore, Principal Co-ordinate analysis (PcoA) of the 16S rRNA gene-based OTUs showed “Treatment” (i.e. control vs. oxic vs. anoxic) to explain over 60% of the observed community diversity among the metagenomes, revealing a strong effect of the prevailing conditions of each treatment.

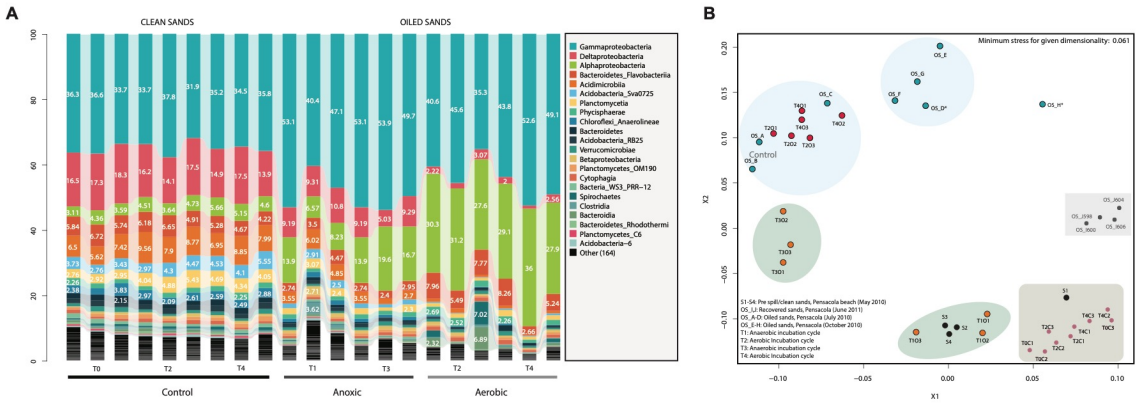


Figure 3-2: A. Microbial community composition shifts during mesocosm incubations. Microbial successional patterns based on 16S rRNA gene-encoding reads. The Y-axis shows the relative abundance (%) of the OTUs assigned to each class. The three samples associated with each time point represent biological triplicates. Note that the controls remain stable through the experimental run. B. Beta diversity shifts between laboratory and field metagenomes. Green and blue circles indicate samples from the anoxic and oxic phases, respectively. Samples with the prefix OS represent field samples collected between May 2010- June 2011, OSA-H represent oil contaminated samples, OS S1-S4 represent pre-spill samples and OS I,J represent recovered samples. T0: baseline data sets at time=0 of the incubation, C1-C3: biological triplicates for the control samples; O1-O3: biological triplicates for the oiled chambers.

3.4.2 Functional gene content shifts reveal rapid microbial community response to oiling

The oxic-anoxic transition had significant effects on the microbial communities of the oiled chambers. These communities were characterized by an initial increase in genes

involved in the biodegradation of simpler aliphatic compounds, followed by increases in genes involved in biotransformation of heavier and more complex/recalcitrant compounds at later stages of the incubation period (Appendix B, Figure B 2). Genes related to nitrogen fixation, nutrient scavenging, stress response and iron acquisition also increased steadily over time, especially during the oxic phases.

Oiling had profound impacts on the genes expressed for nitrogen cycling processes (Figure 3-3, Appendix B, Figure B 6, Table B 1) especially nitrogen fixation, with a substantial increase (>5-fold increase in the metatranscriptome signal) observed the expression of *nif* genes relative to the control. The effect was more pronounced in the oxic phases. This transcriptional response was also consistent with relative abundance assessment at the DNA level; most notably, the number of bacterial genomes harboring a *nifH* gene increased from <5% in the control samples to >65% towards the end of the experiment in the oiled samples (Figure B 5). Interestingly, a large portion (60%) of the relative DNA abundance increase could be attributed to "*Ca.* Macondimonas diazotrophica", which has been previously shown to be a key player in oil contaminated shorelines and a diazotroph (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019). In addition, respiratory nitrate reductases (*narG*) were enriched in the oiled chambers, particularly in the anoxic phases (Appendix B, Figure B 6). Genes for nitrification (specifically ammonia monooxygenases) were below detection in all oiled chambers. The main source of nitrogen inputs into the system appeared to be nitrogen fixation. No genes for anaerobic ammonia oxidation (anammox) were found at the detection limit of our sequencing effort. Oil biodegradation in sediments is often limited by the availability of biologically available (fixed) nitrogen, which is typically exogenously

added during in-situ bioremediation efforts (Pritchard et al. 1992). However, based on our results, nitrogen limitation was alleviated by the selective enrichment of diazotrophs as well as bacteria that catalyze nitrate ammonification/DNRA, which conserves nitrogen in the system, in parallel to substantial and sustained rates of oil biodegradation. The patterns reported above are consistent with available field data (Rodriguez-R et al. 2015; Urakawa et al. 2019). However, in the anoxic phase, sulfate served as the primary electron acceptor based on gene expression patterns of *dsr* genes (dissimilatory sulfite reductase) (Appendix B, Table B 1). Consistent with the gene expression (molecular) data, the sulfate levels in the pore-water reached 5 ± 0.9 mM in the anoxic phase and 14 ± 1.7 mM in the oxic phases indicating the re-oxidation of the reduced sulfur compounds with changes in the oxygen levels in the system. Further, the pore-water nutrient concentrations at the start of the study were ammonium: 6.9 ± 1.2 μM, nitrate: 14.7 ± 2.7 μM, nitrite: below detection and sulfate: 18.0 ± 2.0 mM. The measured salinity value was 35 ppt. The measured pore-water nutrient values during the time course of the experiment are shown in Appendix B, Figure B 6 (lower panel).

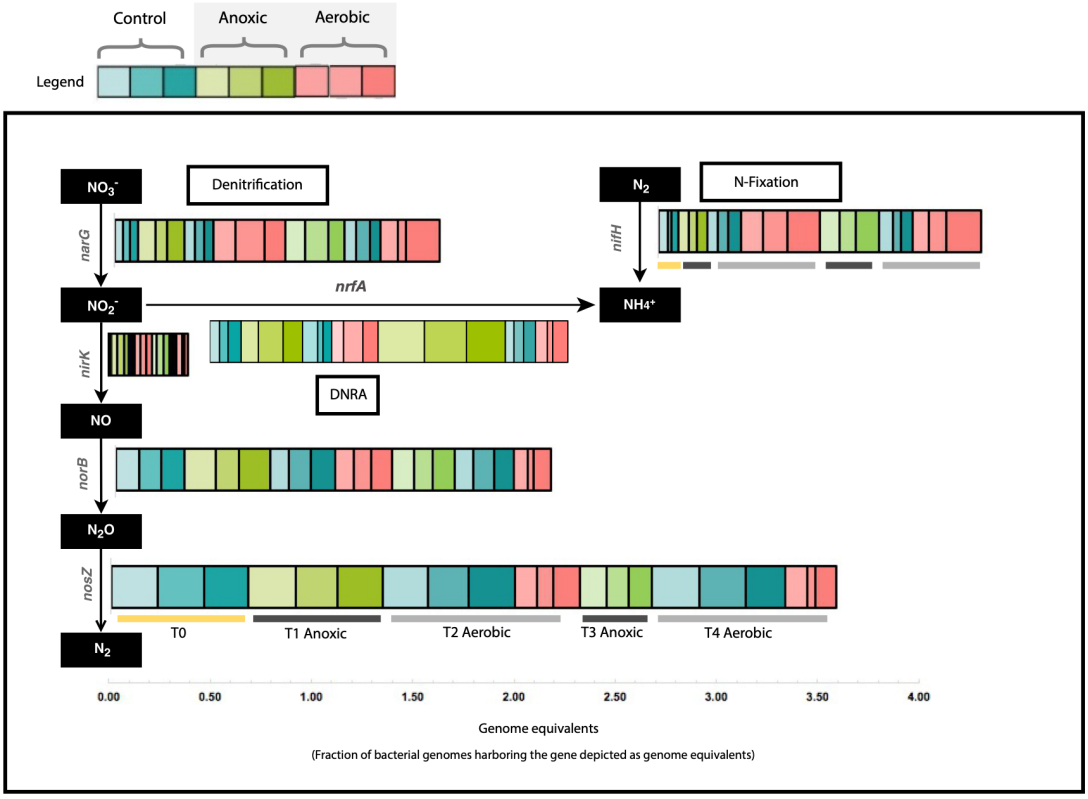


Figure 3-3: Abundance of genetic markers for major nitrogen cycling pathways with incubation time and treatment. Abundance of hallmark genes for denitrification, DNRA and nitrogen fixation pathways the metagenomes (see figure key), represented as genome equivalents (% of total genomes encoding the gene), are shown. The three samples per treatment represent the biological replicates.

3.4.3 Phylogenetic diversity of recovered alkane monooxygenase genes (alkB)

The alkane monooxygenase gene (or *alkB*) has long been established as a reliable biomarker for alkane biodegradation potential. A manually curated reference dataset of 109

alkB genes including putative *alkB* genes identified on contigs assembled from the field beachsand data published by Rodriguez-R et al., 2015 (Rodriguez-R et al. 2015) was used to construct the *alkB* reference phylogeny. The reference dataset spanned a diverse range of taxa and also included sequences recovered from bacterioplankton in the northern Gulf of Mexico (Smith et al. 2013). Subsequently, the metagenomic reads identified as *alkB* were mapped onto the above mentioned reference tree to elucidate their distribution across the *alkB* clades (Figure 3-4). The *alkB* reference phylogeny recovered several distinct clades, with certain clades composed exclusively of (putative) *alkB* sequences recovered from our metagenomic assemblies (Figure 3-4A). These sequences also represented the most divergent clade, underscoring the unexplored diversity of *alkB* that has been elusive to cultivation-based efforts (gray in the inner circle of Figure 3-4A). The read distribution was highly uneven across the recovered clades, with more reads being assigned to a few clades spanning the entire tree, revealing an uneven but phylogenetically unrestrained distribution. Furthermore, distinct *alkB* alleles dominated each oxic phase, presumably linked to the corresponding microbial community successional patterns (i.e., lower length alkanes were degraded earlier/faster followed by the longer ones).

The *alkB*-containing reads primarily originated from control and oxic phase samples as expected since the corresponding protein catalyzes the initial insertion of oxygen. For the anoxic phase samples, the alkyl/methyl succinate synthase (*assA*) gene was used as a proxy of biodegradation activity instead (Appendix B, Figure B 4). The reference tree also included sequences recovered from uncontaminated field beach sands but these clades recruited a relative small number of reads and these reads originated from only form the control chambers (Figure 3-4 B. Panel 3). The results indicated that the

sequences from uncontaminated sites are present at low levels (rare biosphere) in clean sands, and are presumably sustained (selected for) by residual oil or biogenic alkanes.

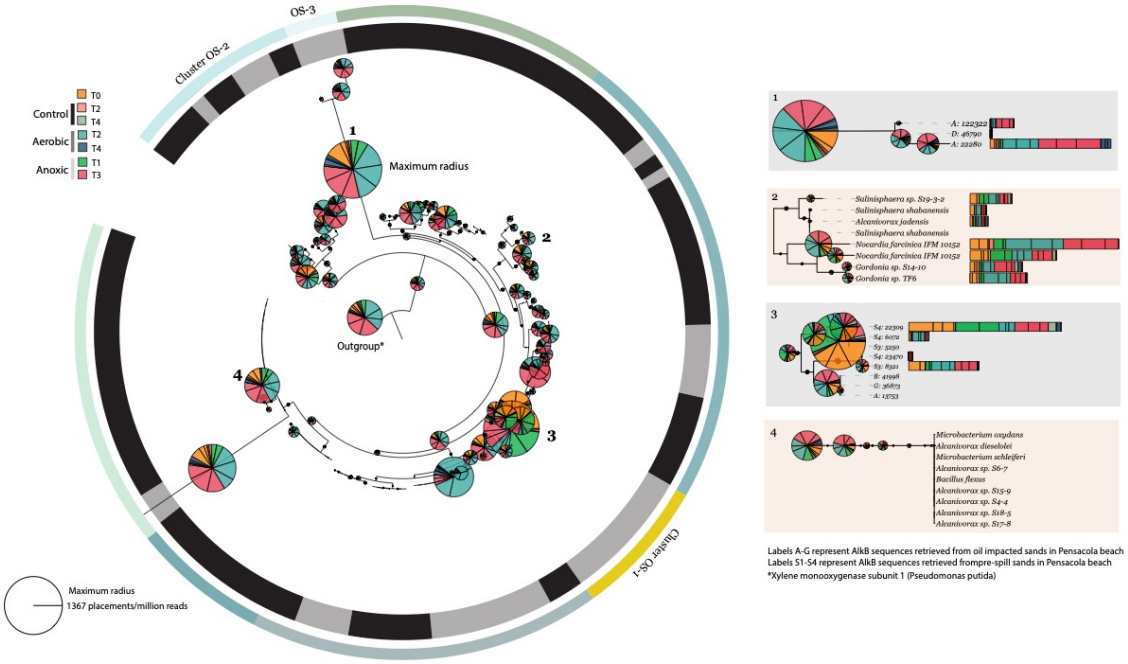


Figure 3-4: Phylogenetic diversity of *AlkB*-containing metagenomic reads. A. The inner circle denotes the origin of the *alkB* reference clade; black: isolates, experimentally verified taxa (based on enzyme assays, gene knockouts etc.), gray: putative *alkB* sequences from metagenome assembly data. Pie chart radii are proportional to the number of reads assigned to the respective clade and the colors represent the treatment (see figure key). The reads have been normalized by the metagenome library size (per million reads). The outer circle denotes the distinct clusters formed by the clades. Clades marked 1-4 have been expanded in Panel B for finer resolution. B. Distribution of the reads in four selected major clusters. 1- Represents the clade that recruited the most metagenomic reads. 2,4-Cluster is composed of verified sequences. Note that only very few reads are recruited by these

clades. 3- Cluster showing the reference sequences recovered form uncontaminated beach sands (labeled S1-S4 on the tree). Similar to the pie charts, the bar charts are proportional to the normalized number of reads recruited by each clade. Xylene monooxygenase subunit 1 (or *xylM*) was used as the outgroup. Detailed version of the phylogeny is provided in Appendix B, Figure B 3.

3.4.4 Genome resolved-metagenomics reveals key metabolic strategies

Co-assembly of the biological replicates at each phase resulted in assemblies of controls with N50 values averaging 1100 bp and those of the oiled samples ranging between 2200-3100bp (Appendix B, Table B 1). The taxonomic classification of the co-assembled contigs as determined through MyTaxa (Luo, Rodriguez-R, and Konstantinidis 2014) is shown in Appendix B, figure B 7. Binning of the assembled contigs into individual population genomes through an iterative binning algorithm (Rodriguez-R et al. 2019) generated 49 metagenome-assembled genomes (MAGs) with completeness >50% and contamination <10%. Twenty-seven of these 49 were high quality MAGs (>80% complete and <5% contamination). At a 95% genome-aggregate average nucleotide identity (gANI), the 49 MAGs represented 31 unique populations (species). Of these MAGs, less than 10% represented known hydrocarbonoclastic or cultured taxa, underscoring the dominance of uncharacterized, uncultured taxa in facilitating microbial community response to oil (Appendix B, Table B 3). Based on genome-aggregate average amino acid identity (AAI) analysis, the previously described genera detected were *Alcanivorax* (62% AAI of MAG to *Alcanivorax dieselolei* strain B5 and 79.59% *Alcanivorax borkumensis* strain SK2).

Between 30% and 50% of the total metagenomic reads mapped on the recovered MAGs, revealing that the MAGs represented a large part of the sampled communities.

Functional gene content and metatranscriptomics analysis of the MAGs revealed a highly dynamic microbial community that adapted to the changing oxygen levels. Specifically, 48% of the 27 high quality MAGs evaluated contained the complete pathway for alkane oxidation, 12% contained the genes for anaerobic alkane metabolism, and 36% and 12% contained the genes for aerobic and anaerobic degradation of aromatic compounds, respectively (Figure 3-5A,B). In addition, 12% of the MAGs possessed genes for dissimilatory sulfate reduction, and found to be more abundant in anoxic phases, presumably sustained by the relatively high sulfate levels (at the mM level) detected in the porewater. Expression of sulfide oxidation genes in the anoxic phase presumably served to prevent toxic sulfide accumulation. Interestingly, ~45% of the MAGs contained the operons for nitrogen fixation, denitrification or DNRA. While MAGs that increased in abundance in both oxic and anoxic phase often contained genes for nitrogen fixation (*nif*), the *nif* gene expression levels were orders of magnitude higher in the oxic phase (Figure 3-5C). Respiratory nitrate reductases were found in MAGs recovered from both the phases (facultative) and their expression was more pronounced only in the anoxic phase (Figure 3-5C). However, sulfate appeared to be the primary electron acceptor in the anoxic stage based on gene expression patterns as well as measurements of sulfate and nitrate concentrations. While more MAGs contained genes for denitrification overall (*nir*; NO_2^- $\rightarrow$NO or "NO forming"), *nrf* genes (formate dependent nitrate reductase, "ammonia forming") were more expressed than *nir* gene in the anoxic phase, presumably reflecting a selection pressure to preserve the fixed nitrogen in the system. None of the MAGs analyzed

here had the genes required for nitrification. This could possibly indicate the higher "sensitivity" of these populations to high levels of organic (oil) loading as hypothesized previously (Rodriguez-R et al. 2015), or that their abundances were too low to reliably assemble their genes and genomes. Sulfate reduction was predominant alternative electron acceptor pathway during the anoxic phase. Interestingly 2/4 MAGs that contained genes for sulfate reduction also contained the complete *nif* operon. All MAGs containing sulfate-reducing genes were affiliated with *Deltaproteobacteria*, consistent with previous knowledge on sulfate reducers (Kleindienst et al. 2014; Meckenstock et al. 2000).

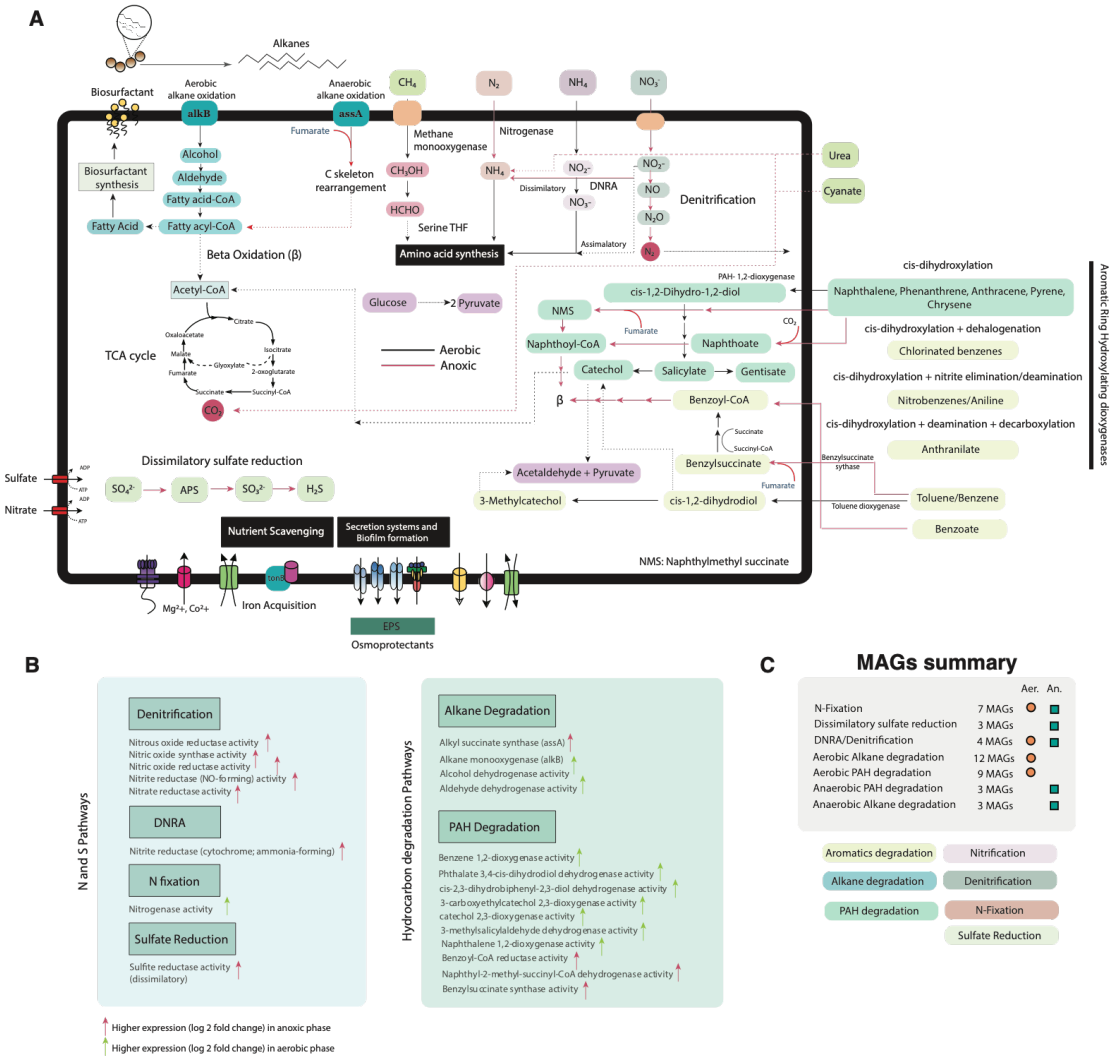


Figure 3-5: . Metabolic pathways present in the MAGs. The key metabolic pathways employed by the MAGs across the two oxygen cycles have been summarized. Red arrows indicate anoxic and black arrows indicate oxic processes. Only high quality MAGs were used here in order to screen for the entire operon. B. Transcriptional responses observed between the two oxygen cycles. Pathways expressed significantly more ($p<0.01$) in either the anoxic or the oxic phase have been indicated by green and red arrows, respectively.

The underlying log2-fold values of gene expression changes are provided in Appendix B, Table B 2, Figure B 6. C. Distribution of predominant pathways in different MAGs.

The MAGs successional patterns followed closely those reported above for individual genes and pathways, whereby the MAGs with genes encoding enzymes for aromatic compounds and the more complex alkanes increased over time and especially during the second aerobic phase (Figure 3-6A). Most notably, the abundance of *nifH* transcripts as well as the genome abundance of "*Ca.* Macondimonas diazotrophica" , a late stage oil degrader increased over two fold, mirroring the patterns observed in our previous field study (Figure 3-7).

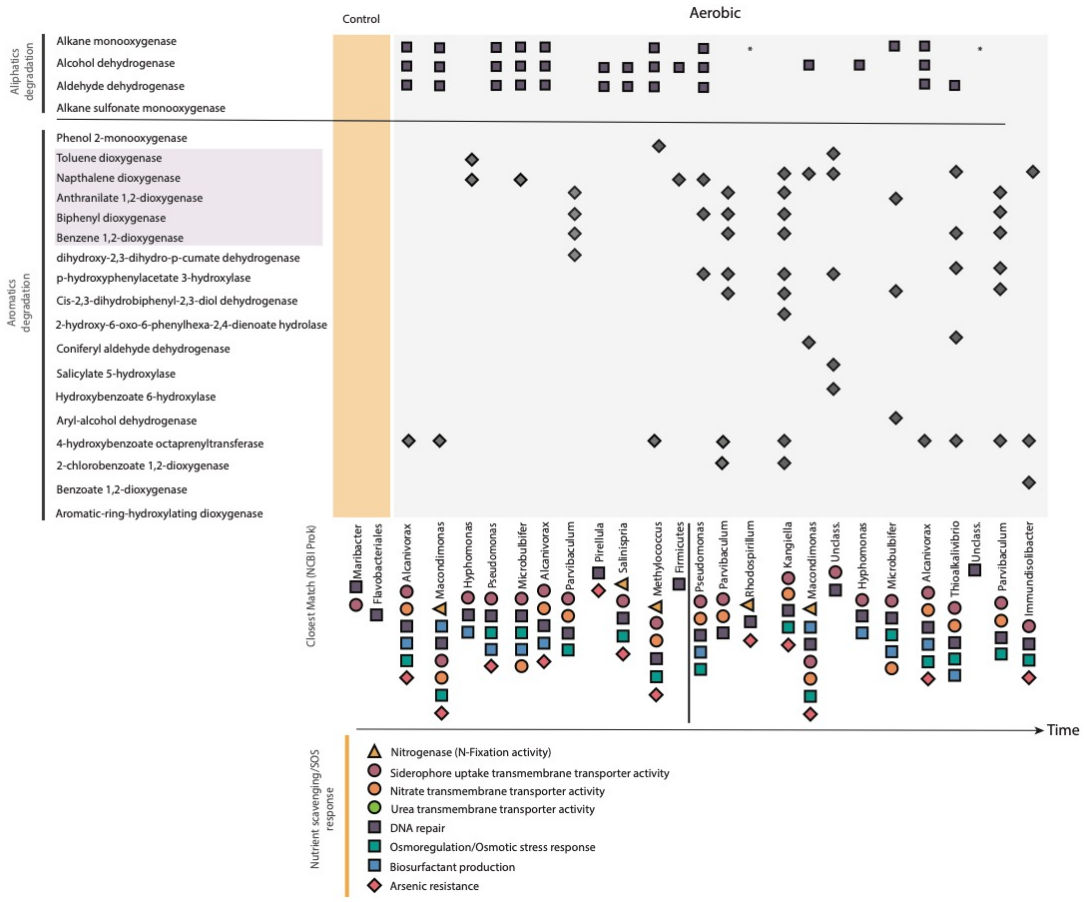


Figure 3-6: Successional patterns of the recovered MAGs indicating substrate specialization. The solid patterns next to the genes indicate the presence of the respective gene (rows) in the MAG (columns). Note the increase in abundance of genes responsible for aromatics degradation.

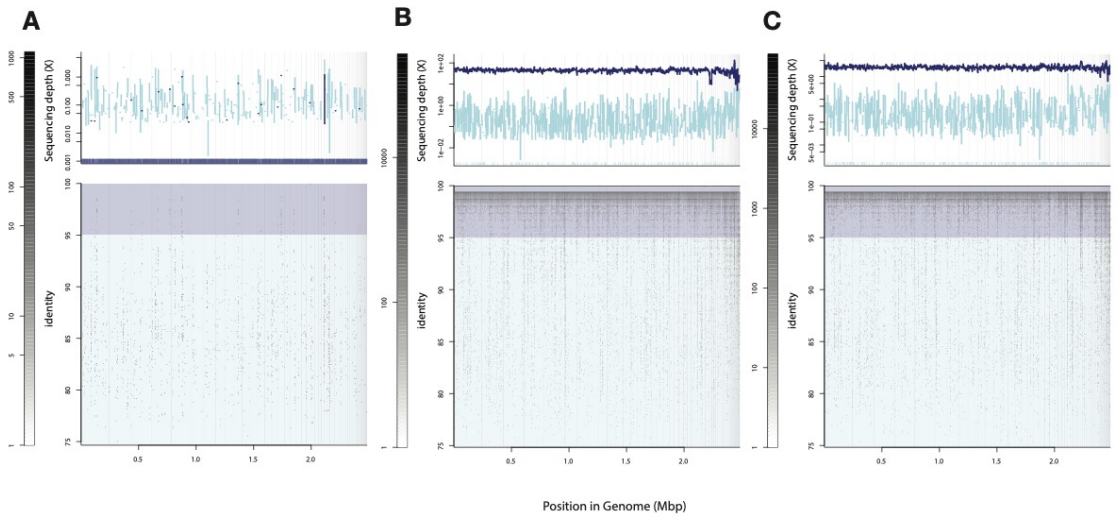


Figure 3-7: Sequence coverage level of *Ca.* Macondimonas diazotrophica in control (A), T2 Aerobic (B) and T4 Aerobic (C) samples. Note the twofold change in abundance from T2 to T4. Read recruitment plot showing the average coverage of the *Ca.* Macondimonas MAG across its genome, in 1,000bp windows, by the metagenomic reads at three time points, i.e., control (A), T2 Aerobic (B) and T4 Aerobic (C). The dark blue histogram 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. The even coverage across the genome reflects a homogenous, sequence-discrete population.

3.4.5 Recovered MAGs as potential biomarkers for ecosystem health

To assess the applicability of our results for *in-situ* processes, the recovered MAGs were searched against available metagenomes from Pensacola Beach (Rodriguez-R et al. 2015). A truncated average sequencing depth (TAD80) was used to reliably detect presence/absence of a MAG in a metagenomic dataset as suggested previously (Tsementzi

et al. 2019). MAGs from the control samples were found in the "Pre-Spill" data set (**Figure 3-8**), revealing that they indeed represent persistent members of clean sands. The MAGs from the aerobic cycle were detected at high abundances in the oiled field metagenomes (i.e., taken after the oil washed up onshore) but the anoxic-enriched MAGs were present at a much lower abundance in these samples. This was consistent with high dissolved oxygen conditions that were measured during field sampling previously (Huettel et al. 2018). However, the detection of anoxic-enriched MAGs in the field samples, even at low abundances, indicated that these organisms may become key players during anoxic periods, similar to what was observed in the our mesocosms. Therefore, the MAGs presented here could potentially serve as biomarkers for the stages of oil biodegradation and ecosystem recovery.

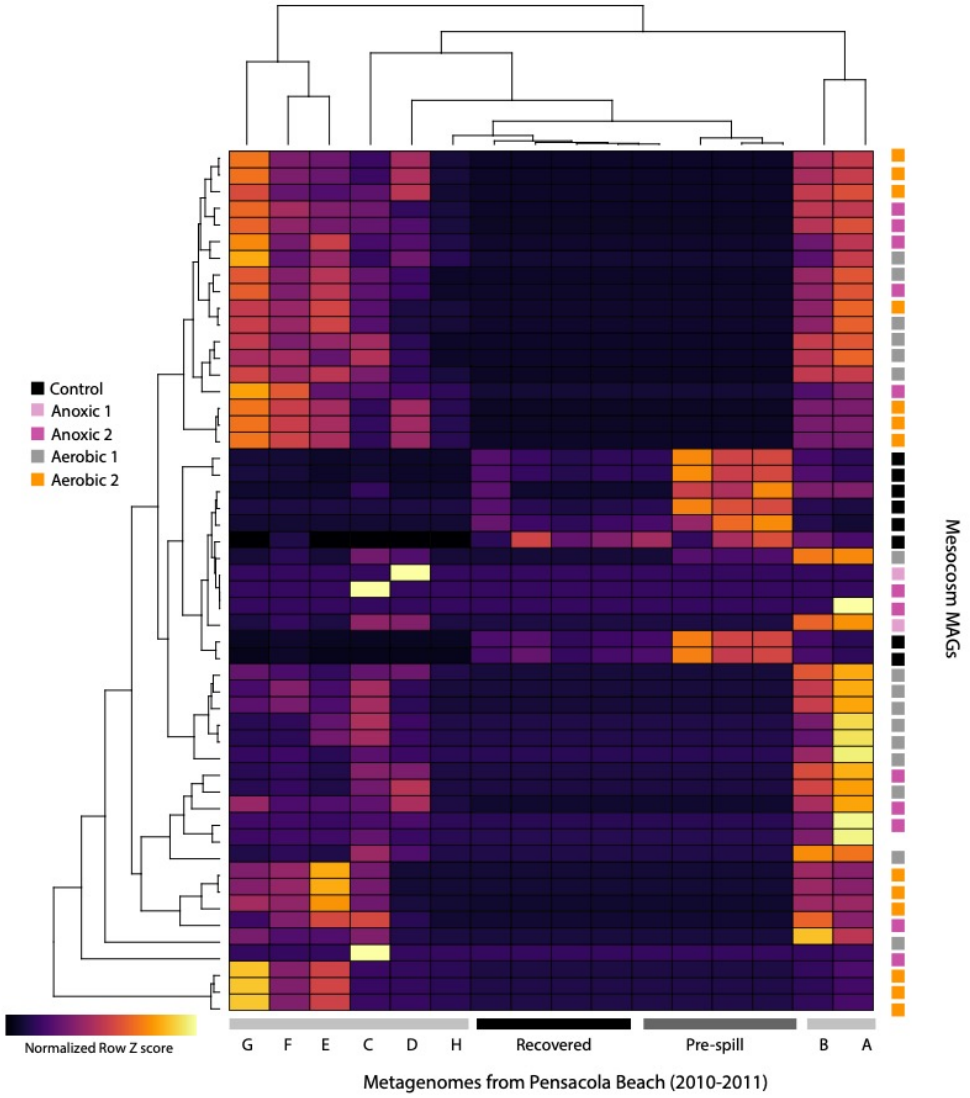


Figure 3-8: Abundance patterns of MAGs in field metagenomic data sets. The graph represents a heatmap of the abundances of each MAG (rows) in each metagenome (columns) estimated as described in the Methods section. Note that the MAGs enriched under anoxic mesocosm conditions were detected at low abundance in the field samples.

3.5 Discussion

Crude oil can persist in sublittoral sediments, tidally wetted intertidal sediments, and dry beach sands, creating a range of conditions for microbial oil biodegradation. The microbial decomposition process is further complicated by the complex composition of the oil, implying similarly complex degradation processes and pathways. Greater than two-fold higher molecular complexity was estimated for buried oil from Pensacola Beach compared to the Macondo crude oil from the DWH wellhead, attributed to microbial and geochemical hydrocarbon modifications and transformations during the degradation process (Ruddy et al. 2014). The transport, distribution and breakdown of petroleum hydrocarbons in marine sediments is also a function of the supply of oxygen, which depends on sediment permeability and the flow rate of gas and water through the sediment pores (Ortega-Calvo and Gschwend 2010; McGenity 2014; Elango et al. 2014; Beazley et al. 2012). Previous studies have shown rates of oil degradation are extremely slow in anaerobic sediments (Shin et al. 2019; Widdel and Rabus 2001b), (Venosa and Zhu 2003), especially in ecosystems like wetlands, saltmarshes and mudflats where the degradation of oil and transport are mass transfer limited (Mortazavi et al. 2013). However, in intertidal sediments, such as the Pensacola Beach studied here, advective transport is considered to influence the rate of oxygen and nutrient exchange (Huettel et al. 2018). The results of this study indicate that oil biodegradation may not be severely limited by the absence of oxygen and may be pronounced in sands where the relatively rapid transport of alternate electron acceptors (e.g., nitrate, sulfate) can support the anaerobic community when oxygen is depleted.

The microbial community shifts studied here revealed microbial communities highly adapted to the fluctuating oxygen conditions with no apparent lag in anoxic oil biodegradation, even after prolonged anoxic conditions that lasted ~18 days. Furthermore, the ratio between oxic and anoxic oil decomposition rates remained fairly stable (Figure 3-1 C), which indicated that the aerobes were not negatively impacted by periods of anoxia and were able to resume oil degradation after oxic conditions were reestablished. Sulfate concentrations in marine sediments are in the mM ranges making them key electron acceptors in coastal marine sediments during periods of anoxia (Jørgensen 1982). Further, the constant ratio between oxic and anoxic decomposition rate, coupled to the detection of the anoxic-enriched MAGs in the rare biosphere of the Pensacola Beach metagenomes, suggested that the intertidal communities are well adapted to dramatic changes in redox conditions and anoxic oil biodegradation.

Frequently, oxygen deprived sediments rich in sulfate accumulate of toxic sulfide due to microbial reduction of sulfate to sulfides, which hamper microbial activity until they are re-oxidized (Broman et al. 2017). In permanently anoxic sediments, the transport of these reduced products to the oxic-anoxic interface could be diffusion limited leading to the accumulation of the toxic sulfides (Burdige 2006). Redox oscillations have been shown to greatly enhance organic carbon remineralization in intertidal sediments when compared to strict anoxic conditions (Burdige 2007). Recently, microbially driven Fenton reactions generated through alternating oxic-anoxic cycles have also been shown to enhance degradation of the recalcitrant compounds (pyrene, anthracene) in oil (Sekar and DiChristina 2017). Reactive oxygen species (ROS) can be generated with Fe(II) in pore waters or Fe(II) produced during the reduction of Fe(III) oxides by dissolved sulfides

(Murphy et al. 2016). The beach sands studied here are characterized by periods of anoxic conditions, followed by frequent oxygenation events due -for instance- to tidal water flow and storms. Interestingly, we found genes for sulfide oxidation as well as sulfide quinone oxidoreductases (SQR) to be expressed during the anoxic phase as well as being present in MAGs enriched in the anoxic phase. SQRs have been previously shown to be a key in marine sulfide oxidation resulting in S^0 or HS_n^- which can further be oxidized via reverse *dsr* pathways (Wasmund, Mußmann, and Loy 2017). This possibly indicated that the beach sand microbes have adapted to the oxygen shifts by oxidizing the generated H_2S as a means of detoxification reflected by the metatranscriptome expression patterns. In addition, benzylsuccinate synthase was highly expressed in anoxic phase indicating the use of these aromatic hydrocarbons as substrate during the anoxic phase with sulfate (or nitrate) as electron acceptor. Collectively, these results revealed, at least in part, the underlying physiology and mechanisms for the observation that microbial oil biodegradation rates may be more enhanced due to redox cycling in these intertidal sediments compared to permanently anoxic sediments with high sulfate levels (Figure 3-1).

In addition to oxygen, nutrients, primarily nitrogen, are known to control the ultimate fate of oil in these ecosystems and be commonly limiting for oil biodegradation (Prince and Atlas 2018; Head, Jones, and Röling 2006; Urakawa et al. 2019). Our study indicated that nitrogen limitation was alleviated primarily by nitrogen fixation (in the aerobic phase) and prevalence for dissimilatory nitrate reduction to ammonium (DNRA), as opposed to denitrification or anammox, in the anoxic phase. Increased abundance of DNRA genes in oil-impacted environments has been noted previously. Low nitrification rates in our mesocosms could be attributed to increased organic matter loading rates that

favor heterotrophs over nitrifiers, which have previously been hypothesize to be more sensitive to oil inputs (Urakawa et al. 2019). Finally, in addition to sulfate, iron can also serve as electron acceptors in marine sediments. The metatranscriptome signals only indicated high expression of genes required for ferrous transmembrane transporters during the anoxic stage and siderophore production in the aerobic phase indicating that the system was iron limited.

The key responder *Ca.* Macondimonas diazotrophica, which made up almost 30% of the microbial community in oiled sands in Pensacola Beach (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019), was detected in similar abundances during the aerobic phases in our incubation study. The detection of almost all recovered MAGs from our laboratory mesocosms in field samples during the DWH incident revealed the usefulness of our closed system for studying the microbial community dynamics that occur *in-situ*. Moreover, screening of the rare biosphere revealed low abundances of the MAGs enriched in the anoxic phase indicating that the field conditions have selected for microbes adapted to fluctuations in oxygen concentrations, in a similar manner to our mesocosm incubations. Metabolic reconstruction of high-quality draft genomes and their abundance patterns generated from this study could potentially serve as biomarkers for elucidating ecosystem health. A phylogenetically unconstrained and diverse set of gene sequences involved in alkane degradation was revealed that belonged to uncultured taxa. These findings further underscored the latent extant diversity that exists in these ecosystems and has only been partially recovered by cultivation or PCR based enumeration studies.

Future studies comparing intertidal communities to supra- and subtidal communities can reveal the key players and important differences and similarities to the

submerged communities studied here. Pensacola Beach experiences diurnal tidal cycles compared to semidiurnal cycles. Currently, experiments employing laboratory systems that simulate the diurnal tidal cycles are underway to explore the microbial community dynamics in temporally saturated coastal sediments.

3.6 Acknowledgements

This research was made possible by grants from The Gulf of Mexico Research Initiative (RFP V Grant No 321611-00 as well as grants to the C-IMAGE II, C-IMAGE III, and Deep-C consortia). Associated NCBI accession numbers have also been provided where applicable. The MAGs and their associated metadata information are available at http://microbial-genomes.org/projects/24. The authors would also like to thank Ioana Bociu from the Department of Earth, Ocean and Atmospheric Science at Florida State University for her help with the hydrocarbon fingerprinting analyses.

CHAPTER 4. A NOVEL, DIVERGENT ALKANE MONOOXYGENASE CLADE INVOLVED IN MARINE CRUDE OIL BIODEGRADATION

Smruthi Karthikeyan, Minjae Kim, Janet K. Hatt, Jim C. Spain, Joel E. Kostka, Konstantinos T. Konstantinidis

4.1 Abstract

Alkanes are ubiquitous in the marine ecosystems and originate from diverse sources ranging from natural oil seeps to anthropogenic inputs and biogenic production by cyanobacteria. Cyanobacterial alkane production in particular represents one of the major sources of C15-C17 compounds in marine systems; accordingly, genes to degrade such short-chain alkanes, e.g., alkane monooxygenase (or *alkB*) are also widespread in marine systems. Crude oil from natural or accidental oil spills is a complex mixture of hydrocarbons, and it remains poorly understood whether known *alkB* are exclusively involved in cyanobacterial and crude oil degradation or, instead, there are novel *alkB* clades specialized in complex, crude oil hydrocarbons. In the present study, large scale analysis of available (meta-) genomic data from the Gulf of Mexico (GoM) oil spill revealed a novel, divergent *alkB* clade recovered from genomes with no cultured/described representatives that was exclusively found in crude-oil impacted ecosystems. By contrast,

the *alkB* clades associated with biotransformation of cyanobacterial alkanes belonged to "canonical" or hydrocarbonoclastic clades. Furthermore, metagenomes of uncontaminated GoM samples, but not those of the global ocean water column samples, revealed the presence of this divergent *alkB* clade, indicating a priming effect of the Gulf for crude oil biodegradation likely driven by the presence of natural oil seeps.

4.2 Introduction

Hundreds of millions of barrels of crude oil enter the marine environment annually as result of natural or anthropogenic activities (Head, Jones, and Röling 2006). Crude oil composition is highly complex, depending on the reservoir source and its underlying biogeochemistry (Overton et al. 2016). However, once the crude oil is released into to the environment, the components undergo rapid changes, which are highly dependent on the *in-situ* physicochemical conditions. (Lea-Smith et al. 2015; Prince, Gramain, and McGenity 2010; Shao and Wang 2013). Following the Deepwater horizon (DWH) accident, massive amounts of crude-oil were released in a short period of time. Alkanes are one of the most abundant components of crude oil, especially in lighter crudes like the DWH oil that contained a higher degree of saturated hydrocarbons. About 50% of DWH oil comprised low to medium length alkanes (Overton et al. 2016; Liu et al. 2012). The site of the DWH incident, the northern GoM (nGoM), is characterized by a plethora of natural oil and methane seeps, whose oil seepage rates are estimated to be over 20,000 m^3/yr. (Orcutt et al. 2010; Teske 2019; Kleindienst, Grim, et al. 2015). This lighter nature of the spilled crude-oil (thus, more inherently amenable to biodegradation) was viewed as the major reason for its rapid consumption by indigenous microorganisms (Atlas and Hazen 2011; Hazen et al. 2010).

In marine systems, widespread biogenic alkane production by algae and cyanobacteria has been documented. It has been estimated that marine algae produce ~308–771 million tons of (predominantly) mid-length, straight-chain hydrocarbons (primarily C15 and C17 alkanes) annually (Lea-Smith et al. 2015; Valentine and Reddy 2015; Schirmer et al. 2010). This estimated value of biogenic alkanes in marine systems far exceeds the input from oil spills and natural seeps (White et al. 2019). Biogenic alkanes are produced via an acyl-acyl carrier mediated decarbonylation of fatty acid metabolites, which also forms the basis for harnessing algae for biofuel (Schirmer et al. 2010). Due to ubiquitous presence of cyanobacteria and algae in surface marine waters, a "latent" hydrocarbon cycle in the ocean has been hypothesized, where these algae-based hydrocarbons could sustain the growth of obligate or facultative hydrocarbon degrading microbes thus, priming the latter organisms to handle oil spills (Lea-Smith et al. 2015). However, crude oil is a more complex mixture of hydrocarbons (including aromatic), and thus, oil-degrading bacteria must also break down complex types of hydrocarbons in oil. It is expected that the ubiquity of algae-produced alkanes in these systems essentially primes the ecosystem to handle exogenous hydrocarbon inputs form other sources like natural or accidental oil releases. It remains currently unclear whether oil-degrading microbes use the same *AlkB* genes for degradation of crude oil as for cyanobacterial hydrocarbons or, instead evolved to develop specialized *AlkB* genes for crude-oil degradation due to presence of natural oil-seeps (Valentine and Reddy 2015).

Alkanes can serve as chemo-attractants to several lineages of bacteria and their active transport mechanisms are hypothesized to determine the chain length that the bacteria have the capacity to degrade (Shao and Wang 2013). Accordingly, the enzymes

responsible for the degradation of alkanes have been broadly classified into short-, mid-, and long-chain alkane degraders. The most frequently encountered ones belong to a class of integral membrane non-heme iron monooxygenases, the alkane monooxygenase or *alkB* which catalyzes the initial hydroxylation of mid-chain alkanes (~C17) (Rojo 2009). The strains carrying this enzyme also frequently carry the soluble cytochrome P450 enzymes. Other enzymes of the hydroxylase family implicated in degrading longer chain alkanes are *AlmA* (Throne-Holst et al. 2007), a putative monooxygenase belonging to the flavin-binding family, and, *LadA,* an oxygenase belonging to the luciferase family (Li et al. 2008). *AlmA* and *LadA* share no apparent homology with the *alkB* or cyt-P450 enzymes. However, to date the genes encoding *alkB* are the most studied due to their ubiquity as well as the availability of experimental evidence for their function. *alkB* and cyt-P450 gene expression if tightly regulated and the transcriptional regulators are frequently found in close proximity to *alkB* and cyt-P450 genes but the details of the regulation and how different alkanes affect gene expression have not been elucidated yet (Shao and Wang 2013).

Incorporation of "omics" based techniques can help to further our understanding of the global metabolic networks as well as the overall process of bacterial alkane-dependent chemotaxis, alkane transport, gene expression regulation and complete mineralization. It remains currently unclear, however, whether oil-degrading microbes use the same *AlkB* genes for degradation of crude oil as for cyanobacterial hydrocarbons or, instead have evolved specialized *AlkB* genes for crude-oil degradation as an effect (selection by) of natural oil-seeps. Genes encoding *AlkB* (alkane hydrolase) are generally considered as a marker for alkane degradation. In spite of its ecological pervasiveness, the reliability of the *alkB* gene as a biomarker for crude-oil biodegradation potential has not been established.

Most of the earlier studies describing *alkB* sequences and their phylogenetic distribution across clades were elucidated using "known" or isolate sequence or through targeted amplicon PCR efforts (Kloos, Munch et al. 2006, Wasmund, Burns et al. 2009, Wang, Wang et al. 2010, Shao and Wang 2013, Smith, Tolar et al. 2013, Nie, Chi et al. 2014). Yet, recent work has shown that the great majority of the microbial populations responding to crude oil (>95% of the total) represent novel taxa not closely related to the isolated and/or characterized ones ((Karthikeyan, Rodriguez-R, Heritier-Robbins, Hatt, et al. 2019); discussed also above). Establishing robust biomarkers using genome-resolved metagenomic approaches can circumvent the biases introduced by conventional PCR or culture-based techniques and offer a much more comprehensive view of functional diversity for ecosystem functioning. Furthermore, establishing robust markers that are specific to crude-oil derived alkanes and not confounded by those involved in the degradation of (freshly produced) cyanobacterial-derived hydrocarbons would greatly help in determining the fate of spilled hydrocarbons in marine systems and assessing, in a more unbiased way, how "primed" these systems are -or not- for (crude) oil degradation.

Following the Deepwater Horizon blowout (DWH), specific taxonomic signatures were observed across spatial and temporal scales accompanied by a short-term decrease in the microbial community diversity in all habitats affected by the spill (Mason et al. 2012; Rodriguez-R et al. 2015; Lamendella et al. 2014; Kimes et al. 2014). Interestingly, the most predominant responders across all the impacted ecosystems belonged to uncultured taxa rather than the previously described hydrocarbonoclastic genera. Previous oligotyping-based studies also indicated that diverse taxa from the rare biosphere dominated the microbial community associated with the deep sea plume (Kleindienst, Grim, et al. 2015).

However, whether these "key" responders are usually members of the rare biosphere in the GoM and in other oceans globally has not been quantified. Assessing the *in-situ* abundances of these organisms with metagenomics would greatly inform whether or not the GoM is more "primed" to deal with sudden anthropogenic oil inputs compared to other, more "pristine" ecosystems. Toward closing these knowledge gaps, we performed large scale analysis of (meta-) genomic data to identify robust microbial indicators or biomarkers for oil biodegradation potential. We screened over 15,000 MAGs spanning a wide range of environments and recovered 615 almost full length *alkB* gene sequences spanning a number of diverse clades. Interestingly, sequences recovered from contaminated nGoM samples were mostly clustered in a novel, deep-branching *alkB* clade, suggesting that GoM microbes harbor distinct *alkB* genes that are apparently tuned for crude oil biodegradation and frequently horizontally exchanged.

4.2.1 Alkane monooxygenase subunit b (AlkB) reference database curation

An initial alkane monooxygenase (*AlkB*) database was constructed using sequences described in (Rodriguez-R et al. 2015) with xylene monooxygenase subunit 1 (*xylM*) gene used as the outgroup. This database was used to retrieve putative full length AlkB protein sequences from assembled scaffolds as well as the MAGs recovered from DWH spill samples. The resulting *AlkB* sequences were manually verified for regions of homology, including verifying the presence of the eight conserved histidine residues (protein active site of AlkB), which are a characteristic motif of *alkB* (Ji et al. 2013). *XylM* was chosen as the outgroup as these hydroxylases have a similar conserved domain though they are not responsible for the oxidation of alkanes (Shanklin and Whittle 2003). An updated *AlkB* database including the newly identified sequences was curated (V2 hereafter). This

database was used to screen for potential *AlkB* sequences contained by the MAGs recovered from a diverse range of environments. Over 15,000 non-redundant, good quality MAGs and over 12,000 isolates encompassing a range of ecosystems spanning bioreactors, natural oil seeps, deep-sea hydrothermal vents, tropical gyres, oil fields and oil tailing ponds, marine water column data from uncontaminated GoM and TARA Oceans expedition (Karsenti et al. 2011) were used.

4.2.2 alkB phylogeny

The *alkB* phylogeny of the sequences retrieved from the above screening (V2) as well as the sequences from the previous (V1) data set were combined and a comprehensive database was generated (V3 hereafter). The V3 sequences were aligned using Clustal Omega v1.2.1 (Sievers and Higgins 2018). The resulting alignment was used for construction of a maximum likelihood phylogenetic tree using FastTree v2.1.7 (Price, Dehal, and Arkin 2009). CD-HIT was used to remove redundancy in sequences (clustered at 95% nucleotide identity), and curate the final *alkB* database, which had representatives spanning all the clades recovered at any point, new and previously described ones (Li and Godzik 2006).

4.2.3 Distribution of alkB short reads in DWH impacted samples

alkB-containing metagenomic short reads were identified using a megablast search against the sequences that made up the reference phylogeny (V2) and subsequently were aligned against the reference alignment using Mafft (Katoh et al. 2002) and placed into the phylogenetic tree using the RaxML EPA algorithm (Stamatakis 2014). The final tree was visualized using iTOL (Letunic and Bork 2016). To build the reference phylogenetic tree

for this analysis, reference sequences were aligned using Clustal Omega (Sievers and Higgins 2018) and the maximum likelihood tree was created using RaxML v 8.0.1 (Stamatakis 2014) using a general time reversible model option, gamma parameter optimization and '-f a' algorithm. The number of *alkB* reads identified from each dataset was normalized by the corresponding metagenome size (reported as number of reads per million reads). *alkB* genes from non GoM datasets were identified with a similar strategy as above.

Genes encoding cytochrome P450 (CYP153) were identified with a similar search strategy to that used for *AlkB* using the reference dataset curated by Wang et al. and Rodriguez et al.(Wang et al. 2010; Rodriguez-R et al. 2015). The reference sequences are available at http://enve-omics.ce.gatech.edu/data/oilspill. Furthermore, genes for *alma* and *ladA* were also identified in all datasets with a similar strategy.

4.2.4 Inferring HGT of alkB genes

To assess putative HGT events, the whole genome phylogeny of the MAGs was compared to the gene phylogeny of the *alkB* genes recovered from the same MAGs. Since alkB is not a single-copy gene, the orthologs for *alkB* had to be identified. This was done using a blast reciprocal best match approach using a combination of -F "m S" -s T options (soft filtering + Smith-Waterman alignment) in blastp as described by Moreno-Hagelsieb and colleagues. (Moreno-Hagelsieb and Latimer 2007). The whole-genome phylogeny was carried out using the corresponding scripts of the enveomics collections as described elsewhere (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019). Dendrograms were constructed for the gene tree and the whole genome phylogenetic tree using the

Dendextend package in R (Galili 2015). The deep-sea plume and the sea-floor sediment communities were considered together due to the considerable overlap of the sampling regions. The generated dendrograms were compared through tanglegrams using the most optimized tanglegram (i.e., least value of entanglement) according to the "step2side" function of the dendextend package.

4.3 Results and Discussion

4.3.1 Phylogenetic diversity of alkB genes and novel clades

To capture as much diversity as this was possible, over 19,800 MAGs and isolate genomes spanning a wide range of environments and clades assemblies from type material of Archaea and Bacteria (as flagged by NCBI) including both complete and draft genomes (n = 12289) were screened for the presence of *alkB* genes. This task resulted in over 600 almost full-length putative sequences that represented 10 distinct clades captured the alkane degradation diversity that exists across ecosystems (Figure 4-1). These clades spanned a wide range of (mostly novel) taxa. The phylogenetic clades represented by *alkB* genes from cultured representatives (Nie et al. 2014; Rodriguez-R et al. 2015) made up a small part of the total clades, e.g., <5% of total *alkB* gene sequences represented isolates, underscoring the predominance of yet-to-culture taxa in *in-situ* hydrocarbon degradation. Interestingly, we found a highly divergent, novel clade of *AlkB* sequences recovered from genomes with no cultured representatives, hereafter termed "DWH clade" (Figure 4-1). The source MAGs of these sequences were all novel genera and recovered exclusively from environments with crude-oil contamination. This clade also included the *AlkB* sequences recovered from the most abundant MAGs in the oiled water column plume,

sediment and beach sand samples affected by the DWH spill (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019; Hazen et al. 2010).

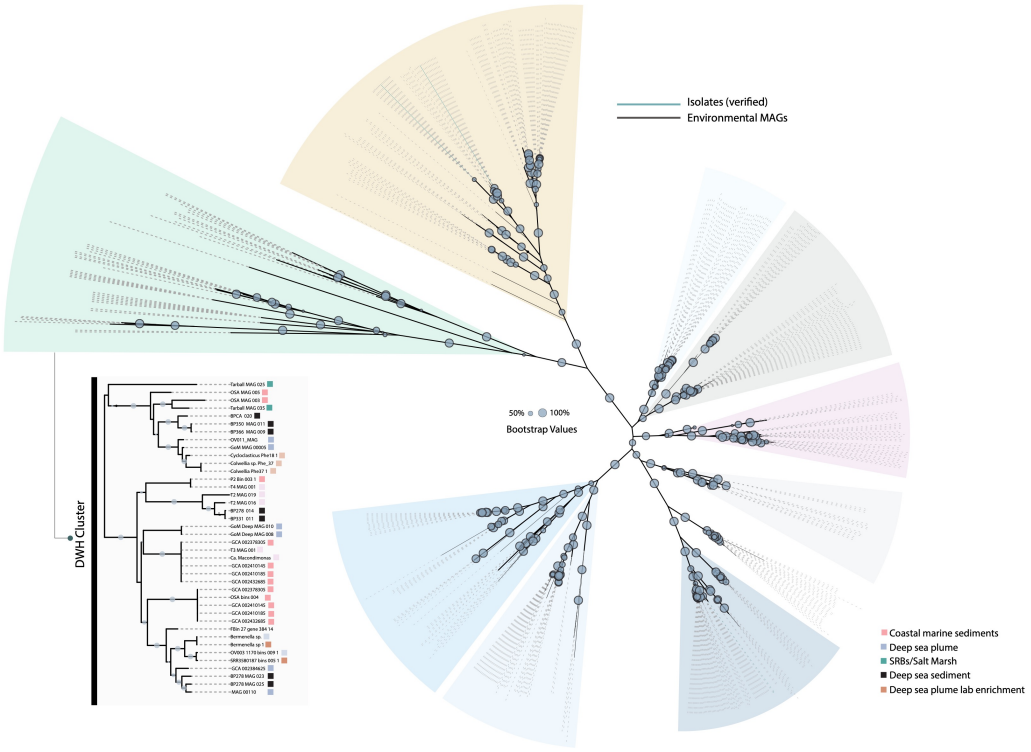


Figure 4-1: Maximum likelihood tree of the full length *AlkB* gene sequences recovered. Sequences were retrieved from screening of >19,800 publicly available MAGs and genomes/draft genomes available in NCBI. The circles indicate the bootstrap values of each branch, only values over 50% are shown. Inset to the left, bottom, shows the environmental context of the sequences representing in the "DWH cluster" (see figure key on the right). Clades indicated by green represented isolates.

4.3.2 Dominance of novel alkB clades in oiled samples

To further study the relative abundance and prevalence of the novel clade revealed, all metagenomic short reads from DWH impacted sites as well as uncontaminated water column samples in the nGoM encoding fragments of *alkB* protein were mapped on onto this reference phylogenetic tree of the ~600 almost full length *alkB* sequences (Figure 4-2). *AlkB* sequences recovered from bacterioplankton in the nGoM reported by Smith et al. (Smith et al. 2013) were also included as a part of the reference phylogeny for this analysis. Over 180,000 reads that met the threshold criteria for a match were mapped onto the reference tree. The majority of the reads recovered by the oil-impacted metagenomes mapped to several distinct clusters (marked "OS" in Figure 4-2). Less than 10% of the reads mapped to clusters made up by "canonical" hydrocarbon degraders or isolate reference sequences, consistent with the genome-based results mentioned above. Furthermore, the recently described *Ca.* Macondimonas diazotrophica (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019) and *Ca.* Bermanella macondoprimitus (Hu et al. 2017) possessed *alkB* genes belonging to this clade.

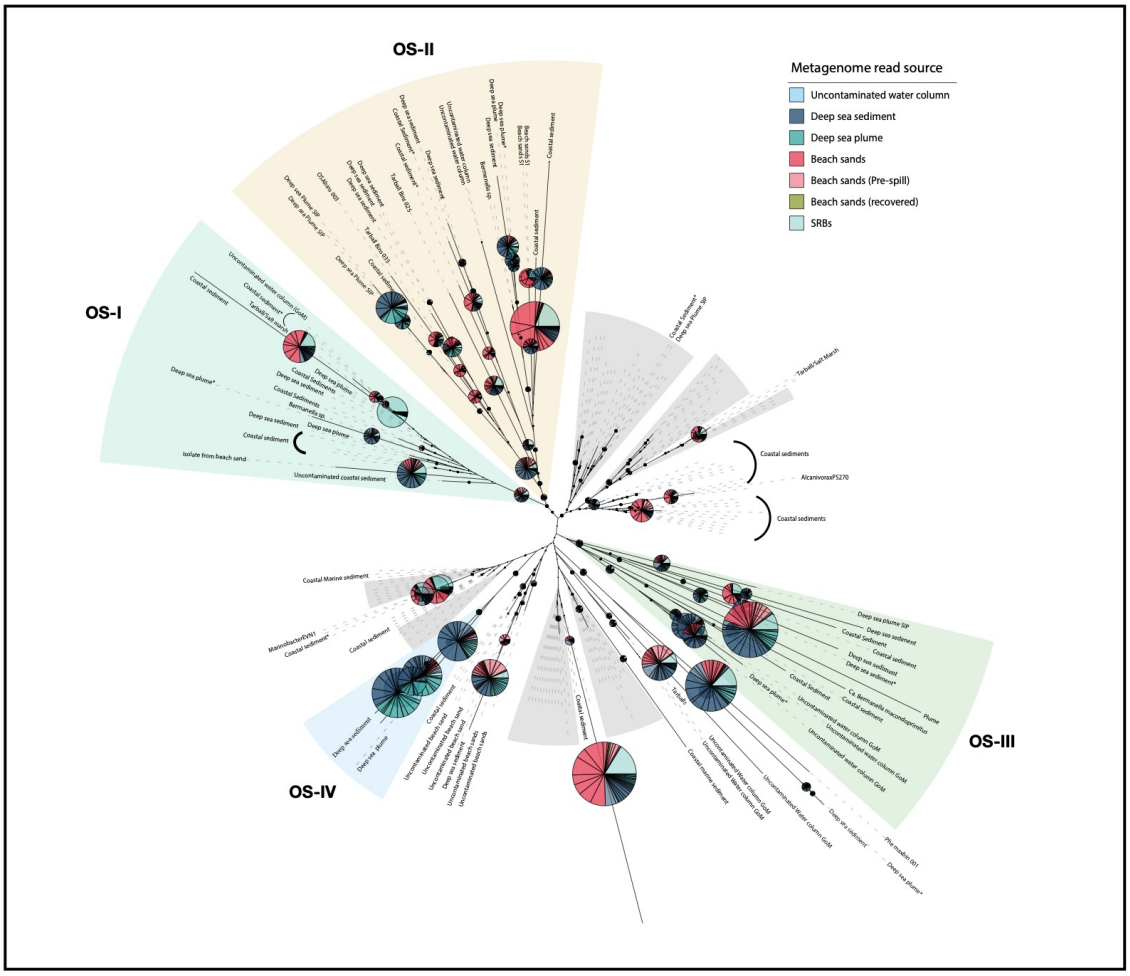


Figure 4-2: Read-based abundance of *AlkB* clades across the sites impacted by the DWH spill. The radii of the pie charts are proportional to the number of reads assigned to each clade (normalized per million reads), and the colors represent the sampling sites from each ecosystem (figure key on top). The clades shaded in grey represent clades that include cultured representatives. Metagenomic reads from contaminated water column (GoM) have also been mapped to these clades to check for their presence in the rare biosphere.

4.3.2.1 *alkB* gene operon organization

The genomic context of *alkB* has been studied previously, and in the ubiquitous and most studied hydrocarbonoclastic bacteria *Alcanivorax borkumensis SK2* genome, it is found in two distinct *alkB* gene clusters/operons (Figure 4-3). The more typical *alkSB$_1$GHJ* operon is implicated in the oxidation of C5-C8 hydrocarbons while the *alkB2* in medium chain alkane degradation. The *alkB2* operon contains a transcriptional negative regulator immediately upstream and a hypothetical protein downstream. In combination, the *alkB1* and *alkB2* operons enable *A. borkumensis* to grow on a range of alkanes up to C32 compounds including branched aliphatic as well as isoprenoid hydrocarbons, alkylarenes and alkylcycloalkanes (Schneiker et al. 2006). Both of these genes, however, were shown to be simultaneously induced in the presence of -alkanes but their underlying regulatory mechanism was not elucidated. In addition to *alkB* genes, *A. borkumensis* also contains cyt-P450 genes which are thought to expand its substrate range. Proteome and qRT-PCR expression data have also shown upregulation of these genes in the presence of mid-chain alkanes as well as isoprenoid hydrocarbons, while genes from *alkB1* or *alkB2* cluster were not expressed at significant levels in the presence of isoprenoid compounds (Schneiker et al. 2006). In the present study, ~ 70% of the cyt-P450 genes detected were present, in addition to the *alkB* genes, in primarily hydrocarbonoclastic clades consistent with previous studies (Wang et al. 2010). More recently, an alkane hydroxylase-rubredoxin fusion gene was cloned from a *Dietzia* strain to *E. coli* and shown to be functional on alkanes up to C40. By comparison, the gene organization of the DWH-clade *alkB* in *Ca.* Macondimonas diazotrophica showed *alkB* gene flanked by genes encoding uncharacterized (hypothetical) proteins. Further studies characterizing the enzymes could

shed light into the regulatory mechanism of this *alkB* gene cluster and its substrate specificity. *Ca.* Macondimonas diazotrophica also possesses multiple copies of the *alkB* gene; however, the organization of those other *alkB* homologs resembled that of the canonical *A. borkumensis*.

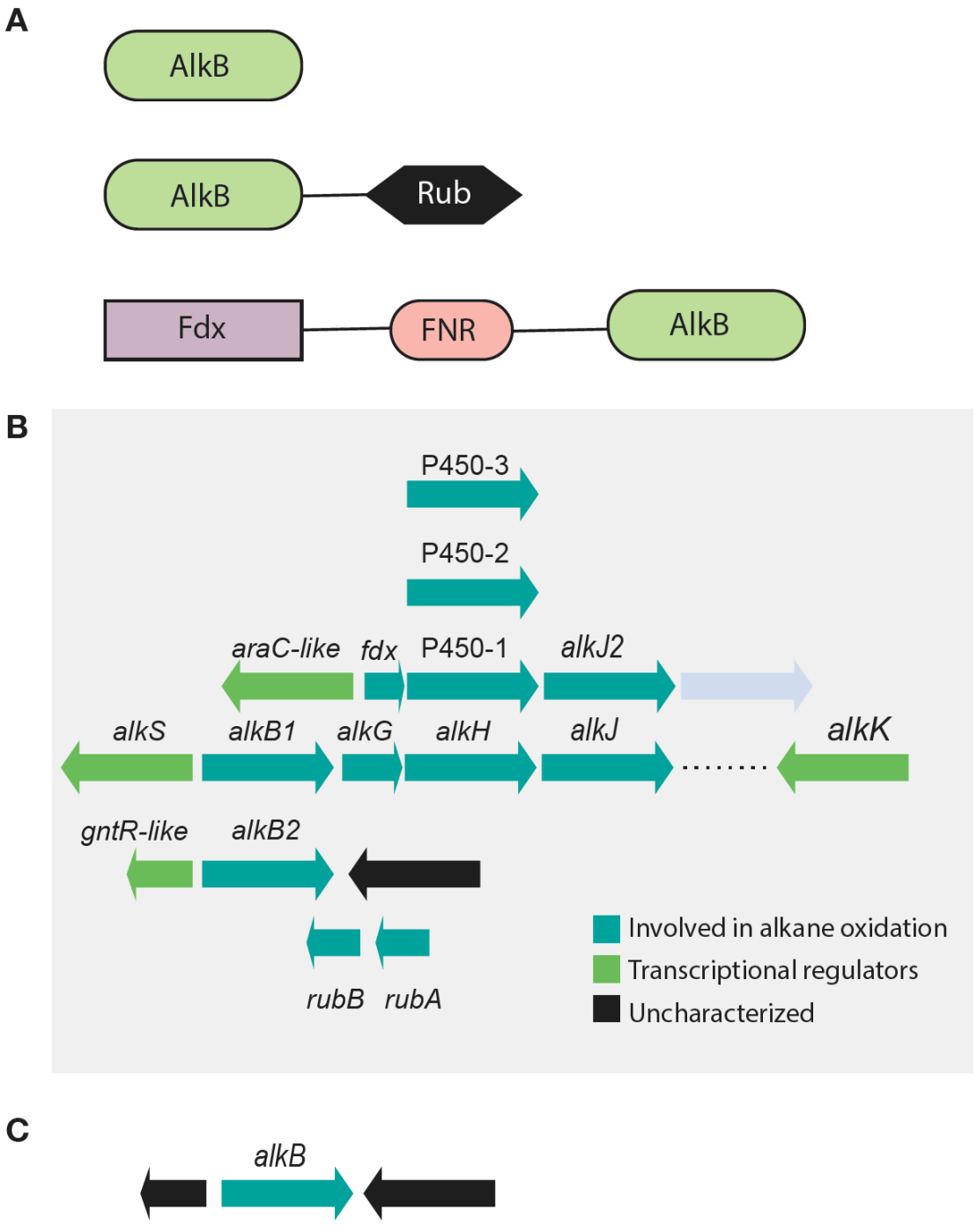


Figure 4-3: *alkB* gene operon organization. Typical organization of the *alkB* gene operon. Rub: rubredoxin, Fdx: ferredoxin, FNR: ferredoxin reductase. B. Organization of

alkB and P450 gene operons in *A. borkumensis SK2*. C. Organization of *alkB* gene operon in *Ca.* Macondimonas diazotrophica (representative of the DWH-clade *alkB*).

4.3.2.2 Expression patterns of the DWH-clade *AlkB*

Available metatranscriptomic data from the field (Mason et al. 2012) or laboratory mesocosm incubation experiments mimicking *in-situ* pressure gradients in beach sands showed consistent patterns, where the genes associated with the DWH-clade *alkB* operon were expressed significantly higher in oiled vs. clean/control samples (log2fold change >2, p-adj=0.01), as well as compared to the non DWH-clade associated reads (Figure 4-4). The other genes namely *almA* and *ladA* implicated in long-chain alkane degradation were not expressed at significantly higher levels. Figure 4-4 shows the expression levels of all the metatranscriptomics reads mapping to *alkB* sequences over the time course of the experiment. The reads with best matches to sequences from the DWH clade are shown in green. The reads assigned to all remaining clades in the reference database are shown in black. The expression of this clade increased with time, possibly indicative of utilization of longer chain alkanes.

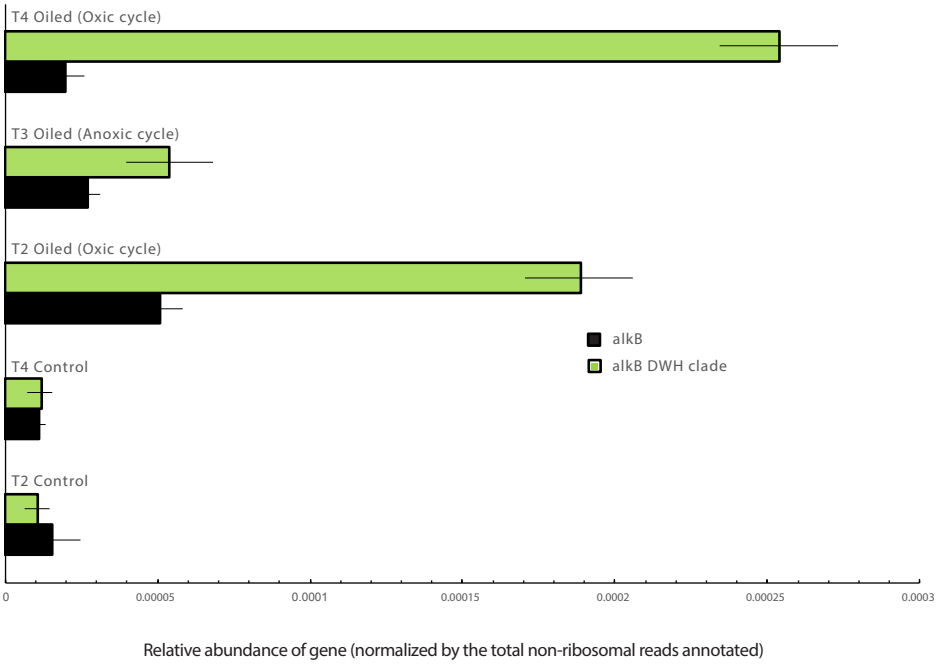


Figure 4-4: DWH clade *AlkB* gene expression patterns in comparison to other alkane degradation genes in laboratory mesocosms. Metatranscriptome analyses showing relative abundance of the annotated reads matching reference sequences from the DWH clade (green) and relative to other *AlkB* clades (black) across the experimental run. Note the comparable baseline levels in the control (no oil added) samples.

4.3.3 alkB prevalence in global marine water column samples

To elucidate whether or not *alkB* clades are specialized to certain habitats, the short-reads identified as containing *alkB* fragments from over 260 marine metagenomic datasets

were mapped onto the *alkB* reference phylogeny (Figure 4-5). A clear trend was observed where the contaminated and uncontaminated GoM samples recruited more of the reads containing the novel *alkB* DWH clade in contrast to the metagenomes from other locations, which recruited more of the *alkB* reads of cultured taxa.

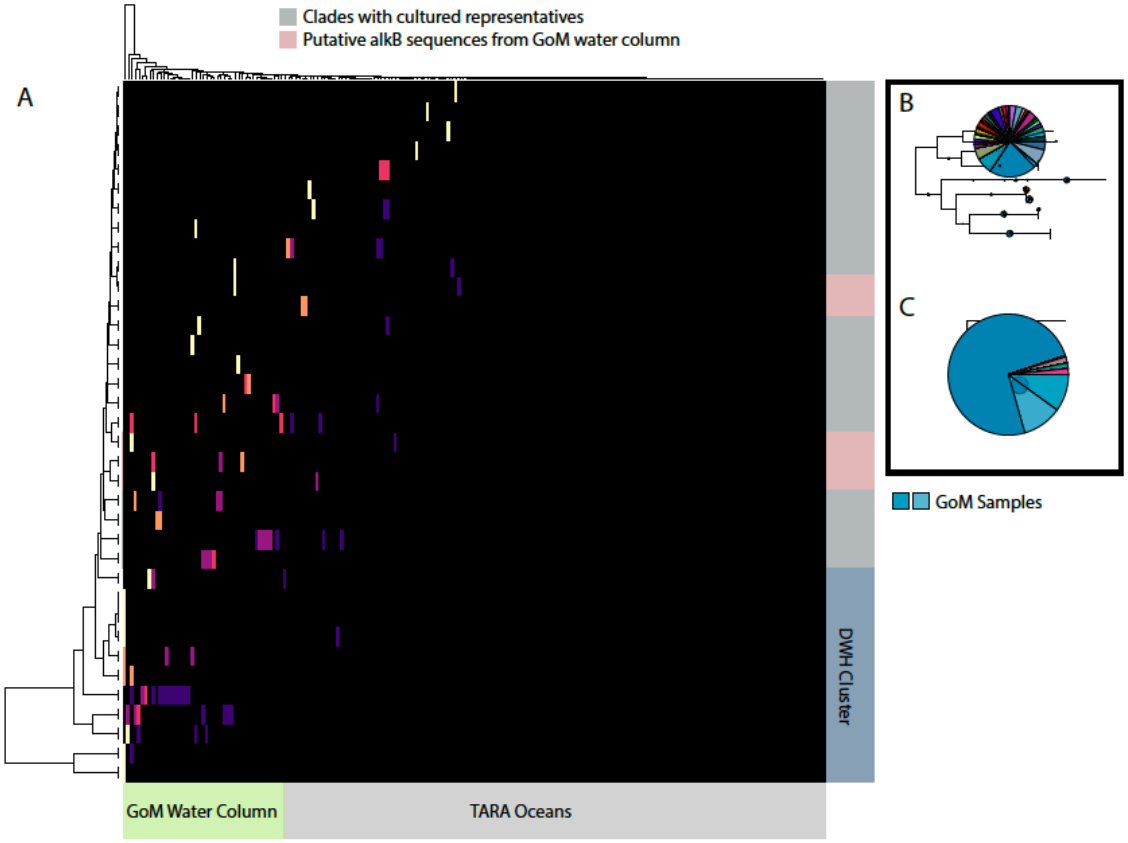


Figure 4-5: Prevalence of *alkB* clades in marine habitats. A. Normalized abundance of each *alkB* reference clade based on reads from 266 metagenomic datasets. Normalization represents number of *alkB* reads per million reads of the corresponding metagenome. Note that sequences from the "DWH clade" are primarily recruited by the GoM samples while the reference sequences associated with known hydrocarbon degraders are more widespread. B. Snapshot of an *alkB* clade composed primarily of known hydrocarbon

degraders (i.e., cultured hydrocarbonoclastic taxa). C. Snapshot of an *alkB* clade from the DWH clade. The distribution of the segments in the pie charts (denoted by different colors) are proportional to the number of reads recruited by the particular reference sequence. The colors in the pie chart correspond to the source of the metagenome data. For instance, the ones shaded in blue were recovered from GoM samples and the rest belong to the TARA oceans sampling dataset.

4.3.4 Horizontal gene transfer of alkB genes

Previous studies assessing the prevalence of *alkB* in various sites, including the nGoM and chronically polluted sub-Antarctic coastal sediments, based on culture-independent qPCR assays and clone libraries also revealed novel *alkB* sequences associated with uncultured genera (Smith et al. 2013; Guibert et al. 2012). However, lack of genomic context of these sequences precluded any further investigation of the frequency of horizontal gene transfer (HGT) events. HGT of *alkB* genes may represent a common adaptive strategy to deal with the distribution of different alkanes in the environment (Guibert et al. 2012; Nie et al. 2014). Consistent with the conclusions of these previous studies, comparison of *alkB* gene phylogeny to that of whole genomes of MAGs encoding the *alkB* gene as part of this study showed high levels of (probable) HGT events (Appendix C, Figure C 1). Moreover, HGT events appeared to be more frequent within habitats vs. between habitats, especially during recent events (i.e. >95% nucleotide identity of the transferred gene) among the two genomes that engage in HGT.

4.3.5 Is the Gulf of Mexico (GoM) "primed" to handle oil spills?

Earlier studies have implicated the GoM's rare microbial biosphere in the rapid degradation of the spilled oil (Hazen et al. 2010; Kleindienst, Grim, et al. 2015; Rodriguez-R et al. 2015). However, none of these have quantified the prevalence of these rare, diverse taxa on a global context in order to assess if the Gulf is more primed for crude oil biodegradation compared to other marine sites due -for instance- to its frequent natural oil seeps. Specifically, we hypothesized that if the Gulf is indeed more primed, then the relative abundance of these primary responders and/or the novel *alkB* clade will be higher in uncontaminated samples from the GoM relative to other marine water column samples. For this, we assessed the abundance of the MAGs recovered from the crude-oil impacted samples against uncontaminated marine water column data from around the globe. An 80% value of the central truncated average sequencing depth (TAD) was used as a proxy to estimate the presence/absence of a certain MAG in a metagenome sample and its relative abundance. MAGs with TAD values > 0.01(or 10% sequencing breadth at 95% confidence) were considered as present as suggested previously (Castro et al. 2018). Based on this TAD threshold values, the abundant MAGs (belonging to uncultured taxa) recovered from the deep sea plume and surface sediments were detectable at the deeper GoM samples (uncontaminated water column) even at low relative abundances, e.g., members of the rare biosphere (Appendix C, Figure C 2). Certain MAGs were detected in TARA oceans data as well, and, upon further inspection, all these MAGs belonged to members of *Alcanivorax spp*., and to a lesser extent*, Pseudomonas spp.* and *Acinetobacter spp.,* all of which had cultured representatives as their close relatives (i.e., the canonical alkane degraders). The MAGs representing the primary responders to the DWH incident were not presence (TAD <0.01 threshold) in these TARA ocean datasets at the limit of detection of the metagenomic

sequencing effort applied to these samples, which was similar to those of the GoM samples. Therefore, the presence of MAGs representing known hydrocarbon degraders in TARA datasets is most likely attributable to the presence of a biogenic hydrocarbons (cyanobacterial origin) in these samples. In contrast, the rapid responders to the DWH spill did not include these "known" oil degraders but rather a bloom of uncultured rare taxa, which are apparently sustained in the deep sea as rare biosphere by the presence of natural oil seeps. (Mason et al. 2012; Kleindienst, Grim, et al. 2015).

Massive next generation sequencing efforts from a wide array of ecological niches impacted by crude oil spills have revealed that >90% of the taxa are yet to be cultured in the laboratory, further underscoring the untapped microbial diversity that exists in nature and responds to oiling. Public repositories host a plethora of sequence data representing these organisms in the form of raw-reads or assembled MAGs from a diverse array of environments. Translating this information into biomarker indicators of the different phases of oil degradation (e.g., early, mid, late) or the nutrient/condition limitation(s) of the microbial communities to perform the biodegradation could be of great importance for site managers and decision makers to help planning responses to oil spills. The divergent, novel *AlkB* clade reported here could be used as such a robust biomarker for evaluating the potential for oil degradation. Studies focusing on substrate range characterization could further shed light on the functional significance this clade. Further studies that measure *in-situ* oil degradation rates and *alkB* expression patterns or stable isotope probing (SIP) experiments tracking the bacteria that assimilate alkane/aromatic compounds are required for more comprehensive assessment of the value of this (and other) clades as a robust biomarker.

CHAPTER 5. GENOMIC EXPLORATION OF DIVERSITY AND ECOLOGY OF OIL DEGRADING MICROBES

Smruthi Karthikeyan, Luis M. Rodriguez-R, Patrick Heritier-Robbins, Janet K. Hatt, Markus Huettel, Joel E. Kostka and Konstantinos T. Konstantinidis

Manuscript under review. Preprint available at: https://doi.org/10.1101/838573

5.1 Abstract

Indigenous microbial communities ultimately control the fate of petroleum hydrocarbons (PHCs) that enters the natural environment through natural seeps or accidental oil spills, but the interactions among microbes and with their chemical environment during oil biodegradation are highly complex and poorly understood. Genome-resolved metagenomics have the potential to help in unraveling these complex interactions. However, the lack of a comprehensive database that integrates existing genomic/metagenomic data from oiled environments with physicochemical parameters known to regulate the fate of PHCs currently limits data analysis and interpretations. Here, we present a curated, comprehensive, and searchable database that documents microbial populations in oiled ecosystems on a global scale, along with underlying physicochemical data, geocoded via GIS to reveal geographic distribution patterns of the populations. Analysis of the ~2,000 metagenome-assembled genomes (MAGs) available in the database revealed strong ecological niche specialization within habitats e.g., specialization to coastal sediments vs. water-column vs. deep-sea sediments. Over 95% of the recovered MAGs

represented novel and uncultured species underscoring the limited representation of cultured organisms from oil-contaminated and oil reservoir ecosystems. The majority of MAGs linked to oiled ecosystems are members of the rare biosphere in non-oiled samples, except for the Gulf of Mexico (GoM) which appears to be primed for oil biodegradation. GROS should facilitate future work toward a more predictive understanding of the microbial taxa and their activities that control the fate of oil spills as well as serve as a model approach for building similar resources for additional environmental processes and omic data of interest.

5.2 Introduction

Oil spills have pronounced impacts on natural ecosystems and the microbial community dynamics following such spills have only recently been documented, as best exemplified by the Deepwater Horizon (DWH) discharge in the GoM, the largest accidental marine oil spill in history. Biodegradation mediated by a complex network of microorganisms is the ultimate fate of the majority of petroleum hydrocarbons (PHCs) that enter the natural environment from accidental discharges (Atlas and Hazen 2011; Hazen et al. 2010; Head, Jones, and Röling 2006; Leahy and Colwell 1990). The microbial interactions that ultimately dictate the fate of oil are nonetheless highly complex and remain poorly understood (Forster, Huettel, and Ziebis 1996). Reconstruction of metagenome-assembled population genomes (MAGs) from metagenomic data (Parks et al. 2017; Hug et al. 2016) enables the recovery of the functional potential of microbial consortia that are associated with critical ecosystem processes such as carbon cycling (Momper et al. 2017; Anantharaman et al. 2016). Consequently, this approach can help to anchor microbial genomes to their biogeochemical functions and ecological niches, and

thus advance understanding of the complex interactions that dictate the ecosystem functions facilitated by microorganisms. Public repositories host a plethora of sequence data from a diverse array of oiled environments, especially raw-reads or assembled MAGs. However, lack of environmental context in the form of *in-situ* physicochemical data that can be easily searched has severely hampered the usefulness and interpretations of the omics data. For instance, a multitude of spatio-temporal "omics" data following the DWH oil spill in the GoM, the first major oil spill for which a lot of omics datasets became available and revealed a short-term decrease in microbial community diversity accompanied by an increase in the functional repertoire (Rodriguez-R et al. 2015). Moreover, specific taxon succession patterns were observed across spatial and temporal scales, which in turn, paralleled the chemical evolution of PHCs (Rodriguez-R et al. 2015; Mason et al. 2012; Kimes et al. 2014). However, the universal applicability of these patterns and the impacts of oil on ecosystem functions as well as the identification of "bioindicator" taxa require further study for use in emergency response efforts.

Specifically, the dynamics of key members of the rare biosphere across the range of environmental parameters observed in oiled marine ecosystems or habitats (e.g., coastal sediments vs. water-column vs. deep-sea sediments) at a global scale requires confirmation with more robust datasets. Such information could validate, for example, universal biomarker taxa and/or genes as indicators for the different stages of oil biodegradation (e.g., taxa that are dominant at early, mid vs. late stages) and thus, help in monitoring the fate of the spills and subsequent ecosystem recovery. Currently, no resource exists to taxonomically classify and provide biogeographic distribution and *in-situ* physicochemical or environmental context for a given MAG to enable these lines of research. The

availability of associated *in-situ* physicochemical data could also help to disentangle the complex interactions that govern microbial community structuring and functioning post disturbance as well as to aid in culture/isolate novel keystone taxa. For instance, our own recent efforts recovered the genome of a population that strongly responded to oil contamination in GoM beach sands that was subsequently shown to comprise 20-30% of the total microbial community in oiled coastal sediments worldwide, and guided the isolation of a representative strain, *Ca.* Macondimonas diazotrophica (Karthikeyan et al., 2019). *Ca*. M. diazotrophica was demonstrated to mediate nitrogen fixation along with hydrocarbon degradation, a combination of traits that likely provide an important ecological advantage in oiled environments that are often nutrient-limited (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019).

5.3 Experimental procedures

5.3.1 Data Curation

All publicly available sequence data at the time of this writing including shotgun metagenome and single-cell as well as isolate genome sequence data associated with the DWH oil spill that includes laboratory enrichments or simulations were used as a part of this study. Metagenome data from uncontaminated water column samples from the GoM were included to elucidate the baseline microbial data and assess the abundance of oil-associated MAGs/SAGs in such samples. To expand this dataset, data from all publicly available crude- or refined-oil contaminated ecosystems including, but not restricted to, natural oil seeps, petroleum reservoirs, coal fields, hydrocarbon contaminated sites, and methane gas vents were also included. All metagenomes and MAGs and SAGs were

obtained from the NCBI or MG-RAST databases (Keegan, Glass, and Meyer 2016). Details on the data sources are provided in Supplemental Table 1.

5.3.2 Quality Control and Trimming

The raw metagenome reads from the publicly available data (single or paired end) were trimmed using Trimmomatic (dynamic trim option) (Bolger, Lohse, and Usadel 2014) and quality checked using the SolexaQA (Cox, Peterson, and Biggs 2010) package with a cutoff of Q>20 (>99% accuracy per base-position) and a minimum trimmed length of 50 bp. For certain metagenome datasets that had shorter average sequencing read lengths (~75 bp), a minimum read length of 30 bp was used as the threshold.

5.3.3 Assembly and Binning

Assembly and binning were carried out for the metagenomes that did not have associated MAG data publicly available. Assembly (or co-assembly for metagenomes similar in composition) for the single-cell and metagenomic datasets was performed using IDBA-UD with default parameters unless otherwise noted (Peng et al. 2012) and only contigs >1000 bp were retained for binning. For time- or spatial-series metagenomic datasets (mainly sediment samples), an iterative subtractive binning pipeline was employed to enhance recovery of high-quality MAGs. In order to determine which of the metagenome sample sets could be co-assembled for the iterative binning, MASH, a tool employing the MinHash dimensionality reduction technique (Ondov et al. 2016) was used to evaluate the pairwise distances between the metagenomic datasets. The resulting distance matrix was then visualized using NMDS (non-metric multidimensional scaling) plots. A combination of MASH distances and a Markov Cluster Algorithm (MCL) using

the script “ogs.mcl.rb” from the Enveomics toolkit (Rodriguez-R and Konstantinidis 2016b) that identifies Orthology Groups (Ogs) in Reciprocal Best Matches (RBM), was employed to evaluate the samples that could be pooled together for co-assembly; samples with distance values below 0.1 were co-assembled. For the sediment metagenome samples alone, the paired end reads were merged using PEAR (Zhang et al. 2014) with –p 0.0001 setting. The resulting reads (non-merged forward, non-merged reverse and merged) were combined and quality trimmed as described in the earlier section and sequences less than 70 bp were discarded. IDBA-UD with the following parameters: --long, --mink 25, maxk 121, --step 4 was used to assemble the resulting reads. These parameters provided longer contigs as well as improved recovery rates of MAGs from these environments.

Population genome binning was carried out using MaxBin v2.0 (Wu, Simmons, and Singer 2016) and the MAG’s completeness and contamination were assessed via CheckM (Parks et al. 2015). MAG quality was determined as [Completeness – 5*(Contamination)]. All MAGs with quality score >50 were used in downstream analyses. High quality MAGs were defined as Completeness > 75% and Contamination < 5%, and medium quality MAGs were defined by Completeness >50% and Contamination > 10%. To determine redundancy of the MAGs (when multiple methods were used to recover MAGs, MAGs were de-replicated using a 95% ANI level over at least 20% of the genome using FastANI v1.1 (Jain et al. 2018).

5.3.4 Spatio-temporal abundance distribution of DWH impacted MAGs

In order to assess the biogeography of the DWH impacted MAGs within and across habitats, the short reads from contaminated or non-contaminated metagenomes were

mapped onto the MAGs. For this, the MAG sequences were initially tagged and combined into a single Bowtie 2 database (Langmead and Salzberg 2012). The metagenomic reads were then mapped onto this database using a competitive search using Bowtie 2. The 80% central truncated average of the sequencing depth or the TAD80 value was estimated for the mapped metagenome reads as described elsewhere (Tsementzi et al. 2019). The estimated sequencing depth was normalized using the sequencing depth of a single-copy gene *rpoB* as described in Tsementzi et al. 2019 (Tsementzi et al. 2019), which provide the normalzied relative *in-situ* abundance estimate across datasets. Normalized MAG abundances were visualized using a Heatmap generated using the heatmap3 package in R (v3.4.0) (Zhao et al. 2014).

5.3.5 Taxonomic and functional annotation of MAGs

Genes were predicted for the MAGs using MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010) and annotated using the curated SwissProt database as described elsewhere (Johnston et al. 2016). The resulting annotations were filtered by their amino acid identity and alignment score and bitscore using as a minimum match >= 40% AAI, >= 70% alignment length, bitscore >= 60. The SwissProt database identifiers were mapped to their corresponding metabolic function based on the hierarchical classification subsystems (Level 1 of the SEED subsystem category) (Overbeek et al. 2014) in order to provide broad functional category information. Genes relevant to oil biodegradation were manually verified. Heatmaps depicting the functional gene content across samples were generated using the heatmap3 package in R (v3.4.0). The MAGs were taxonomically characterized based on whole genome AAI and ANI comparisons against the NCBI prokaryotic database using MiGA (Rodriguez et al. 2018).

5.3.6 Whole genome phylogeny of the oil-associated MAGs and SAGs

Genes were predicted for all of the 1864 high quality MAGs recovered from crude-oil contaminated environments using Prodigal v2.6.1 (Hyatt et al. 2010). Only the MAGs with over 75% completeness were used in the whole-genome phylogeny reconstruction. A set of 106 universally present "marker genes" were identified in these MAGs using the script 'HMM.essential.rb' available as a part of the enveomics toolkit. The resulting reads were then aligned using Clustal Omega v1.2.1 and concatenated using the script 'Aln.cat.rb', which removes any invariable sites. The phylogenetic tree was constructed using RaxML using the PROTGAMMAAUTO option (1000 bootstraps). The resulting tree was visualized using iTOL (Letunic and Bork 2016).

5.3.7 Generation of interactive maps

Maps used in the webserver were created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license.

5.4 Results and discussion

In the present study, we curated the " **Genome Repository of Oiled Systems**" ("GROS"), a comprehensive MAG and single-amplified genome (SAG) reference database that is provided as an independent and searchable project through the Microbial Genomes Atlas webserver (MiGA) (Rodriguez et al. 2018). Since the DWH discharge represents the first major oil spill occurring after the development of next generation sequencing technologies, GROS is currently dominated by DWH data as well as MAGs binned from

publicly available data from other oil-impacted environments, such as natural oil seeps and lab incubation or enrichment studies designed to mimic oil spills under near *in situ* conditions (Supplemental Table 1). The interactive graphical interface allows for browsing the MAGs/SAGs/isolate genomes in the GROS project and their associated metadata. GROS also enables users to query their isolate or MAG genome sequence (partial or complete) against its reference MAGs and SAGs to identify if the queried genome represents a new taxon (species) or is another member of taxa that is already represented among the reference MAGs. For query genomes of the same species as a reference MAG, the GROS project can be used to assess the global biogeographical distribution of the reference MAG on an interactive map and obtain information about the habitats/samples where the latter has been recovered and its relative abundance within these samples (when a corresponding metagenome is available). Therefore, the resulting data can help determine the taxonomic uniqueness of a queried genome, its relative *in-situ* abundance, and the extent of association with oiled samples based on existing data. The taxonomic classification of reference or user-provided query genomes is performed by MiGA as described previously based on the ANI/AAI concept (Rodriguez et al. 2018). Incorporation of embedded GIS-based "layers" as part of the metadata description of each reference MAG provides an interactive interface that enables filtering based on criteria including location, habitat, taxonomy, date, nutrient concentrations, and hydrocarbon data, among other parameters (Figure 5-1). The incorporation of the MAGs and their associated metadata as GIS layers enables overlaying additional *in-situ* physicochemical data available through the Environmental Response Management Application, ERMA

(https://response.restoration.noaa.gov/gulf-mexico-erma), an online mapping tool which contains real-time data collected from the GoM.

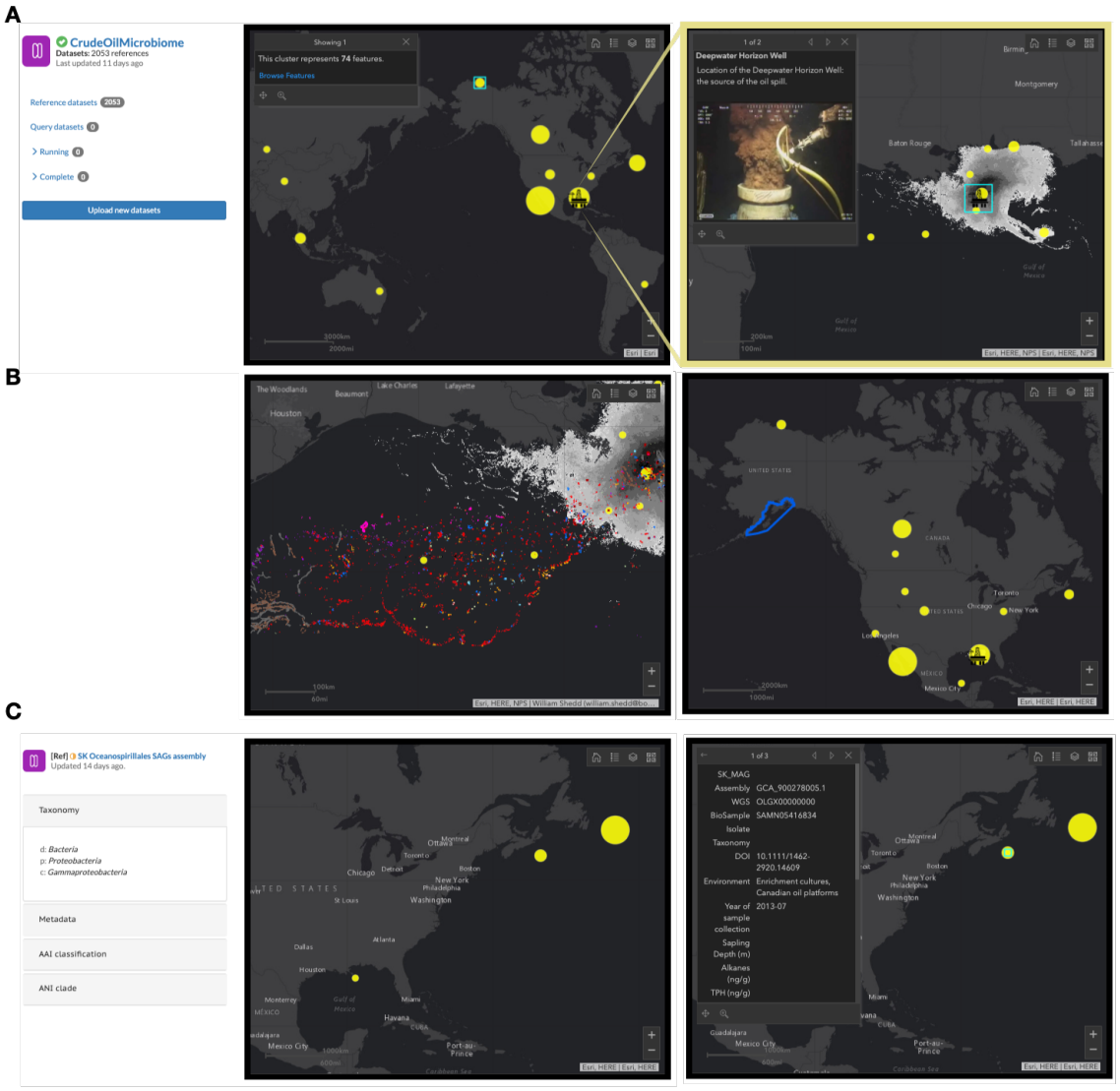


Figure 5-1: The graphical output from the GROS webserver. A. The map shows the distribution of all MAGs present in the curated database. The circles are proportional to the abundance of MAGs associated with each location. The top right panel shows only the MAGs recovered from the Gulf of Mexico (GoM) and the grey area denotes the cumulative

days of oiling following the DWH accident along with the location of the Macondo MC252 wellhead. The interactive map allows enabling/disabling any or all of the other layers; namely, oil spill boundaries from previous oil spills, natural seep locations based on seismic surveys, as well as filtering the MAGs based on user-defined genomic criteria such as taxonomic name and genome completeness level. B. Enabling additional layers. The map layer showing the natural seep locations in the GoM region has been enabled. Bottom right: The modified map when the layer containing previous oil spill boundaries (in this instance, the Exxon-Valdez oil-spill) have been included. C. Assessing MAG biogeography. When a query genome is searched against the GROS database, the closest MAG match in the database when at an ANI level >95% (same genomospecies) along with the physicochemical data of the sample from which the MAG was recovered and relative abundance in other samples will be shown on the interactive map and the left panel, respectively. The specific example shown represents the distribution of an unclassified *Oceanospirallales* MAG detected in the deep-sea plume during the DWH oil-spill.

The reference database consists of 2021 MAGs and SAGs, 1864 of which are of high quality (i.e., >75% completeness, <5% contamination) that were previously made available or were recovered from available metagenomes using established assembly and binning techniques(Bowers et al. 2017) as part of this study. For the latter, an iterative binning methodology (Tsementzi et al. 2019) was employed in certain cases where multiple metagenomes from a time or spatial series were available to recover additional high quality MAGs (mainly for the DWH impacted sediments); otherwise individual metagenomes

were assembled and binned. The MAGs are grouped into species-like clusters based on the 95% genome-aggregate average nucleotide identity (gANI) threshold as recommended previously and implemented in MiGA (Konstantinidis and Tiedje 2005), and only one representative genome for each cluster is used when a genome is queried against the database in order to reduce the CPU demand. A total of 1536 unique clusters (inter-cluster ANI < 95%) were recovered among these MAGs, including 22 clades with 5 or more members, thus revealing extensive species diversity for microbial taxa associated with oiled environments. Cultured taxa were observed in (only) 4.25% of all the recovered 95% ANI clusters (Figure 5-2). The 2021 MAGs were assignable to 63 different bacterial and archaeal phyla based on genome-aggregate amino acid identity (gAAI) values (Figure 5-2), with over 40 of them affiliated with Candidate Phyla Radiation (CPR) as well as the newly described Asgard superphylum (Seitz et al. 2019; Zaremba-Niedzwiedzka et al. 2017). Recent studies have shown that Asgard archaea have the metabolic potential to carry out the anaerobic oxidation of short chain hydrocarbons (Seitz et al. 2019). Moreover, notably, nearly 45% of the DWH MAGs had AAI values between 40-50% to the closest described species with available (isolate) genome(s), which represents at the least novel families (Rodriguez et al. 2018) and further underscores the novel microbial diversity that exists in these oiled niches (Figure 5-3).

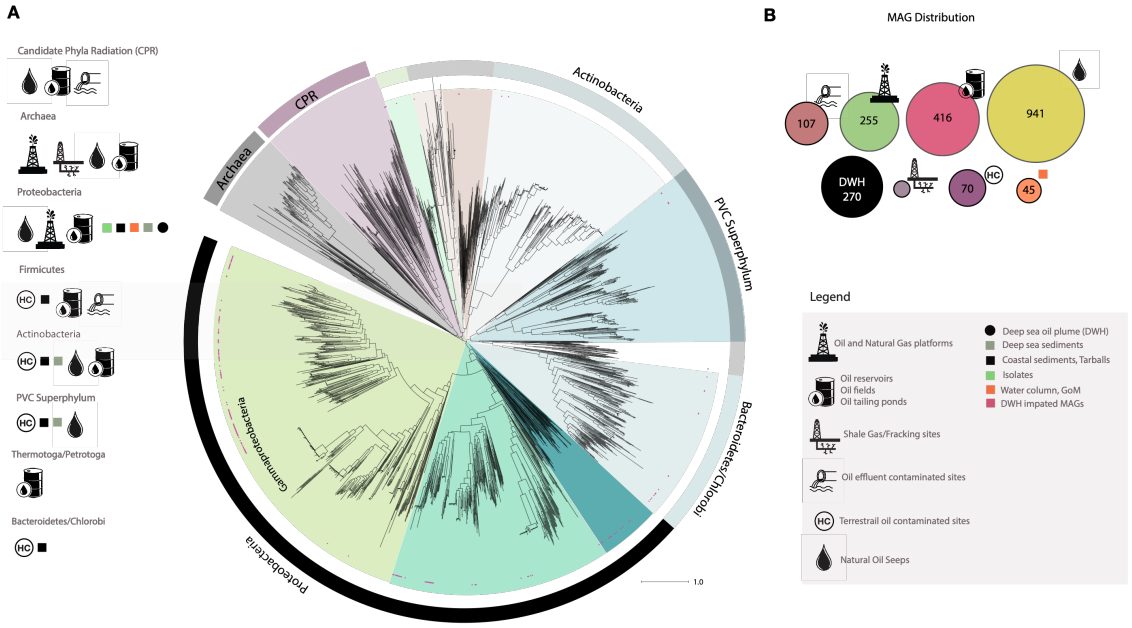


Figure 5-2: Whole-genome phylogeny of the oil-associated MAGs. A. The universal core gene phylogenetic tree of the1864 high-quality, oil-associated MAGs, when overlaid with the habitat of origin of the MAGs, reveals strong ecological niche specialization. For instance, proteobacterial MAGs appear widespread and not localized to a certain niche. However, MAGs from the CPR and Asgard superphyla were only associated with oil reservoirs, tailing ponds and deep-sea sediment in the vicinity of natural seeps. Similarly, MAGs from shale or natural gas fracking sites were either archaea or belonged to *Firmicutes* (specifically, the recently described "*Ca*. Frackibacter"). MAGs from *Thermotoga* or *Petrotoga* (extremophiles) were exclusively found in oil reservoirs or oil wells. Note that most of the MAGs recovered from DWH impacted sites (indicated as red dots in the tree) were assignable to *Proteobacteria* or *Bacteroidetes*. B. Total MAGs recovered from each ecological niche. Ecological niches are defined in the figure legend.

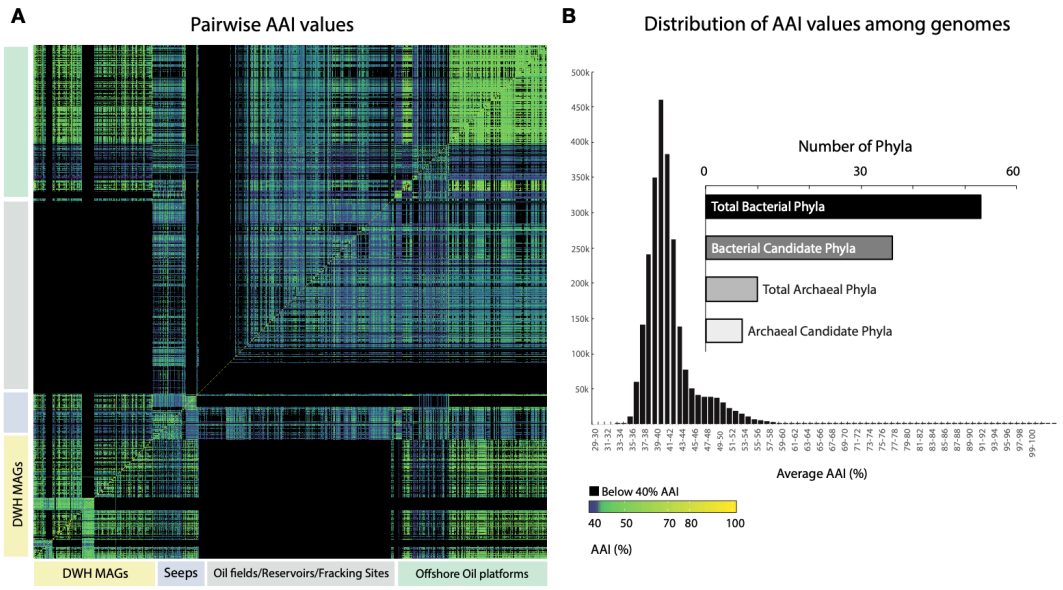


Figure 5-3: Genetic relatedness of the genomes in GROS. A. Pairwise genetic relatedness among MAGs based on AAI. Black boxes represent AAI values below 40%, blue boxes represent AAI values between 40-45%. B. Distribution of AAI values among the MAGs. Inset: Total number of bacterial and archaeal phyla represented by these MAGs.

The universal gene phylogeny revealed a strong ecological niche specialization for the majority of the MAGs and rather limited geographical distribution with a few exceptions. For instance, 82 % of the high quality DWH MAGs and SAGs were strictly associated with specific habitats of the Gulf, i.e., deep sea surface and subsurface sediments, deep sea oil plume or sediments and surface residue balls (or SRBs) recovered from coastal ecosystems, and they rarely crossed between these distinct habitats. Interestingly, at the 95% ANI (or species level), lab incubation studies that aimed to mimic

the *in-situ* conditions after the DWH spill shared between ~18% and over 60% of the MAGs recovered from *in-situ* samples for the deep-sea plume (Hu et al. 2017) and coastal beach sands (Karthikeyan, Rodriguez-R, Heritier-Robbins, Kim, et al. 2019), respectively, which represents a much higher fraction compared to similar laboratory incubations for other microbial processes (Shin et al. 2019; Bacosa et al. 2018). At higher taxonomic levels (i.e., genus or higher), more overlap was observed between the lab simulations and respective field data (>80% overlap at the genus level) as well as between distinct habitats (~30% overlap) affected by the DWH spill.

MAGs associated with the DWH spill (across all habitats) predominantly belonged to the *Proteobacteria* phylum whereas a higher proportion of MAGs from sediments associated with natural oil seepage environments, petroleum reservoirs, and shale gas/fracking sites belonged to the recently proposed Candidate phyla radiation or CPR (Hug et al. 2016). Furthermore, archaeal MAGs were only found in the vicinity of natural deep-sea seeps, oil reservoirs and shale gas/fracking sites and were not detected in any of the DWH impacted samples. This is likely attributable to the lack of dissolved oxygen and the presence of gaseous hydrocarbons at the latter sites when compared to the DWH-impacted sites, most of which were aerobic. Archaea are typically associated with the presence of methane and other short-chain and gaseous hydrocarbons as is the case in natural oil seeps where dissolved oxygen is often not detectable (Seitz et al. 2019). In contrast, the deep sea DWH plume at the seafloor was aerobic, presumably leading to enrichment of taxa that differed from those in the sub-surface samples. Further, the early stage composition of the deep sea DWH plume was rich in gaseous hydrocarbons, and as previous reports suggested, these gaseous compounds were instrumental in priming the

microbial populations for breakdown of the more complex substrates of the plume (Valentine et al. 2010). GROS can help in advancing such discoveries by documenting the abundances of MAGs along with the associated hydrocarbon and physicochemical gradient data and providing insights into the biodegradation process based on the MAGs' successional patterns.

Mapping of metagenomic reads to the recovered MAGs was employed to determine the degree of biogeography (i.e., limited distribution globally) of the MAGs. For this analysis, reads from each metagenome were mapped against the MAGs at high nucleotide stringency (>95% nucleotide identity) and a non-zero TAD80 (*i.e.*, Truncated Average sequencing Depth across the genome after removing the top and bottom 10% genome positions in terms of sequencing depth) was used as the threshold to determine presence vs. absence of a MAG in the metagenome. Note that non-zero TAD80 corresponds to at least 10% sequencing breadth, corresponding to a confident detection (presence) at species-level resolution (ANI ≥ 95%) as suggested previously (Castro et al. 2018). This abundance information is also available through GROS for the reference MAGs of each 95% ANI cluster. In general, MAGs showed strong biogeography with the notable exception of the deep-sea oil plume and sea floor MAGs that showed broad distribution across geographic distances, even at the species level. More specifically, ~16% and ~17% of the MAGs recovered from the deep-sea oil plume and sediments belonging primarily to class *Deltaproteobacteria* and *Dehalococcoidia* were detected (i.e., same species) in the metagenomes recovered from the natural seeps (GoM, Santa Barbara Channel, Guaymas Basin) and seawater in the vicinity of oil platforms (Canada), respectively. There was considerable overlap between the beach sand and SRB communities across distinct

geographic locations in the GoM (Figure 5-4 A-C). Further, less than 10% of the coastal sediment associated MAGs (mostly proteobacterial) were observed in terrestrial crude-oil impacted metagenomes and typically at low abundances (Figure 5-4 D). Finally, the global marine metagenome data of the TARA Oceans expedition (exclusively uncontaminated samples) were only recruited by the uncontaminated GoM MAGs and, to a lesser extent, by the genomes of known hydrocarbonoclastic bacteria namely, *Alcanivorax spp*. and *Acinetobacter spp*. This finding could indicate that these hydrocarbon-degrading bacteria maintain substantial (not rare) populations in-situ based on growth on biogenic alkanes of algal or cyanobacterial origin, which are prevalent in the open, non-oil-contaminated ocean (Lea-Smith et al. 2015). In contrast, at least 40% of the MAGs recovered from the DWH oil plume and seafloor sediment impacted by the oil spill were not detected in the TARA ocean datasets (rare biosphere) but were found in low abundances in the metagenomes from the uncontaminated water column in the GoM. Collectively, these results could indicate that certain populations respond to oiling, i.e., increase from undetectable or low abundance to represent a substantial faction of the microbial community, typically >0.5-1% of the total upon oiling and are part of the rare biosphere in uncontaminated sites, but tend to be more abundant in the GoM due to abundant natural seeps in this region. In other words, the GoM appears to be more "primed" for crude-oil biodegradation based on our preliminary data.

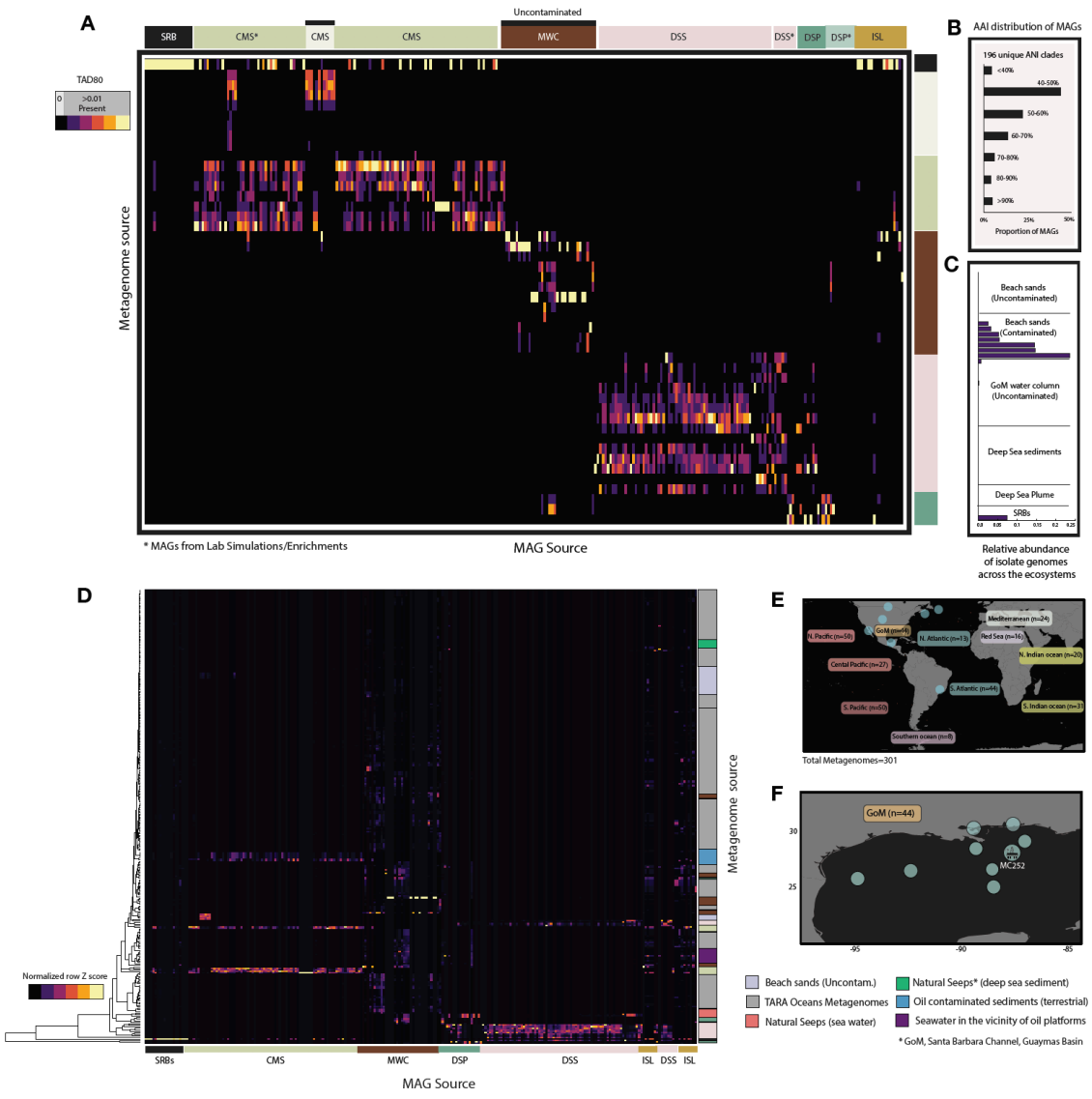
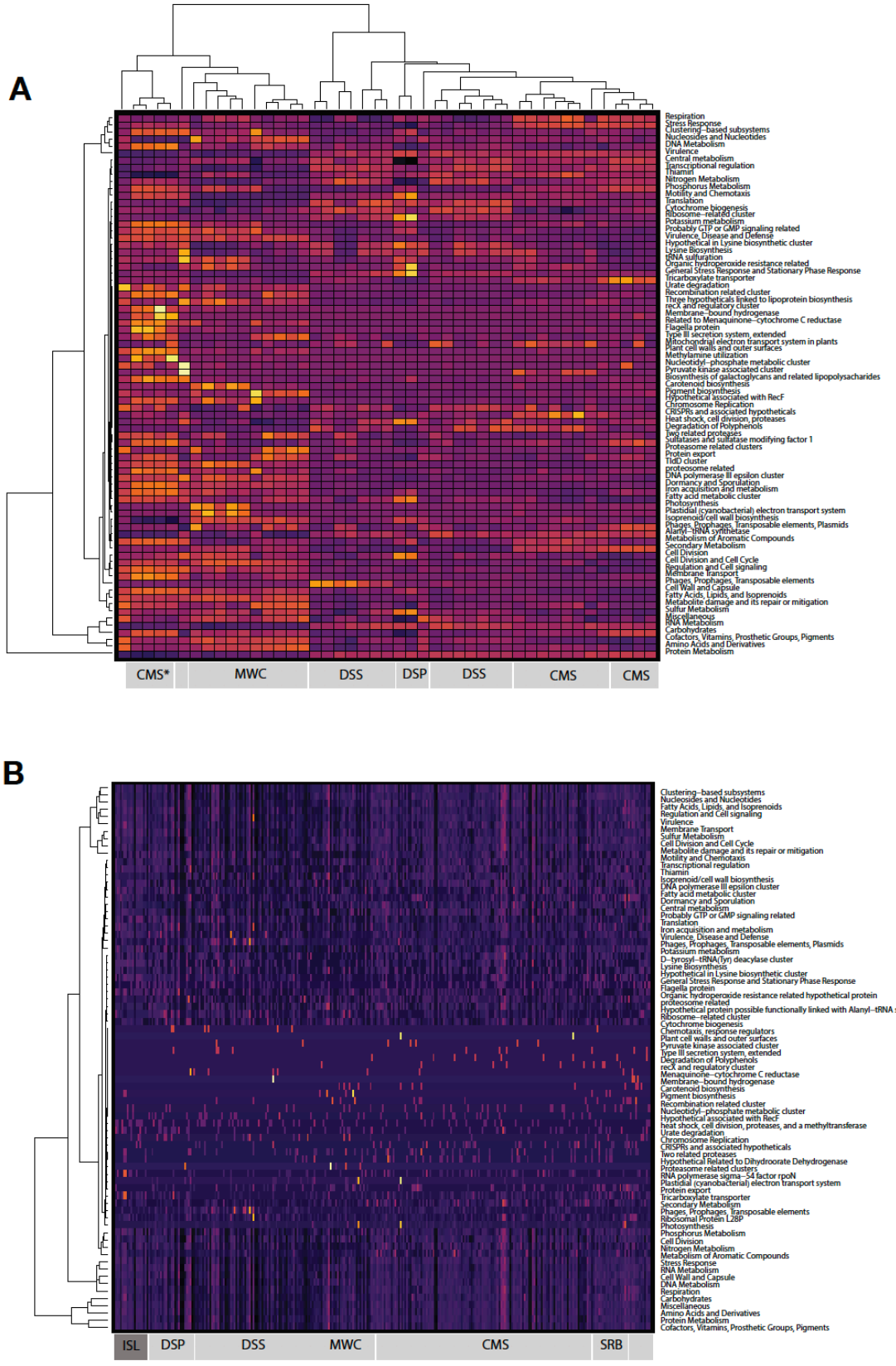


Figure 5-4: Biogeographic distribution of the DWH-associated MAGs. A: Heatmap showing the TAD80 values of DWH-associated MAGs in various metagenomes from the GoM (top x-axis); absence of a MAG in a metagenomic dataset (y-axis) is indicated in black color. S1-S4 represent lab simulation or enrichment data; uncontaminated water column samples were also included for comparison. Note the minimal overlap that occurs between habitats with the exception of the deep-sea sediment and plume. Substantial overlap also exists between the beach sand and SRB communities. B. Distribution of AAI

values between the DWH associated MAGs and their closest described relative in the public databases showing that ~50% of the MAGs represent novel family or higher taxonomic ranks (e.g., best match AAI <50%). C. Proportion of reads from the metagenomic datasets mapping to isolate genomes of oil degraders showing that less than 5% of the reads mapped to these genomes. The notable exception is beach sands and SRBs where the recently described isolate "*Ca.* Macondimonas" made up >25% of the total community after the oil spill. The relative abundance for each dataset was normalized by their corresponding *rpoB* gene sequencing depth as described in the Methods section. D. Similar to Panel A but including 301 non-GoM metagenomes from uncontaminated global water column (includes TARA Oceans data), deep-sea natural seeps data from the GoM, the Santa Barbara channel and Guaymas Basin, crude-oil contaminated sediments and seawater data in the vicinity of oil and natural gas platforms, and oiled terrestrial (soil) metagenomes. Note that some of the uncontaminated water column samples shown (brown for GoM, grey for others) have very low or undetectable abundances for most of the oil MAGs, reflecting that they are members of the rare biosphere especially in non-GoM samples. E. Map showing the locations of the metagenomes used. F. Similar to panel E but showing only the GoM sampling locations. For simplicity, the habitats have been broadly categorized as Deep-sea sediment (DSS), Deep-sea oil plume (DSP), Coastal marine sediments (CMS), Surface residue balls (SRB) and marine water column (MWC) for uncontaminated samples recovered from the GoM.

In contrast to the strong biogeographical patterns largely observed at the individual MAG level, functional annotation of the metagenome reads and the MAGs associated with DWH impacted regions showed much less separation by habitat at hierarchical or subsystem levels (Figure 5-5 A), i.e. specific functions did not show biogeographical or habitat-specific associations but the genomes that encoded them did. At a 95% ANI (or 95% nucleotide identity for individual genes), the DWH MAGs were grouped into 195 unique clusters whose MAG composition was highly dependent on the source habitat of the MAGs. However, when the proteins predicted from these MAGs were clustered at a 40% amino acid identity (subsystem level classification), no such clear clustering pattern was observed (Figure 5-5 A, B). Moreover, the proteins involved in oil degradation (specifically, alkane monooxygenase as well as ring-hydroxylating dioxygenases involved in degradation of the aromatic hydrocarbons) did not show habitat preference indicating that these functions may have moved horizontally in the recent or more distant past between habitats (Figure 5-5 A,C). Incomplete lineage sorting and secondary gene losses could also be other possible mechanisms underlying this observed pattern (Maddison and Knowles 2006).


A
Respiration
Stress Response
Clustering-based subsystems
Nucleosides and Nucleotides
DNA Metabolism
Virulence
Central metabolism
Transcriptional regulation
Thiamin
Nitrogen Metabolism
Phosphorus Metabolism
Motility and Chemotaxis
Translation
Cytochrome biogenesis
Ribosome-related cluster
Potassium metabolism
Probably GTP or GMP signaling related
Virulence, Disease and Defense
Hypothetical in Lysine biosynthetic cluster
Lysine Biosynthesis
tRNA sulfuration
Organic hydroperoxide resistance related
General Stress Response and Stationary Phase Response
Tricarboxylate transporter
Urate degradation
Recombination related cluster
Three hypotheticals linked to lipoprotein biosynthesis
recA and regulatory cluster
Membrane-bound hydrogenase
Related to Menaquinone--cytochrome C reductase
Flagella protein
Type III secretion system, extended
Mitochondrial electron transport system in plants
Plant cell walls and outer surfaces
Methylamine utilization
Nucleotidyl-phosphate metabolic cluster
Pyruvate kinase associated cluster
Biosynthesis of galactoglycans and related lipopolysacharides
Carotenoid biosynthesis
Pigment biosynthesis
Hypothetical associated with RecF
Chromosome Replication
CRISPRs and associated hypotheticals
Heat shock, cell division, proteases
Degradation of Polyphenols
Two related proteases
Sulfatases and sulfatase modifying factor 1
Proteasome related clusters
Protein export
TldD cluster
proteosome related
DNA polymerase III epsilon cluster
Dormancy and Sporulation
Iron acquisition and metabolism
Fatty acid metabolic cluster
Photosynthesis
Plastidial (cyanobacterial) electron transport system
Isoprenoid/cell wall biosynthesis
Phages, Prophages, Transposable elements, Plasmids
Alanyl-tRNA synthetase
Metabolism of Aromatic Compounds
Secondary Metabolism
Cell Division
Cell Division and Cell Cycle
Regulation and Cell signaling
Membrane Transport
Phages, Prophages, Transposable elements
Cell Wall and Capsule
Fatty Acids, Lipids, and Isoprenoids
Metabolite damage and its repair or mitigation
Sulfur Metabolism
Miscellaneous
RNA Metabolism
Carbohydrates
Cofactors, Vitamins, Prosthetic Groups, Pigments
Amino Acids and Derivatives
Protein Metabolism
CMS*
MWC
DSS
DSP
DSS
CMS
CMS
B
Clustering-based subsystems
Nucleosides and Nucleotides
Fatty Acids, Lipids, and Isoprenoids
Regulation and Cell signaling
Virulence
Membrane Transport
Sulfur Metabolism
Cell Division and Cell Cycle
Metabolite damage and its repair or mitigation
Motility and Chemotaxis
Transcriptional regulation
Thiamin
Isoprenoid/cell wall biosynthesis
DNA polymerase III epsilon cluster
Fatty acid metabolic cluster
Dormancy and Sporulation
Central metabolism
Probably GTP or GMP signaling related
Translation
Iron acquisition and metabolism
Virulence, Disease and Defense
Phages, Prophages, Transposable elements, Plasmids
Potassium metabolism
D-tyrosyl-tRNA(Tyr) deacylase cluster
Lysine Biosynthesis
Hypothetical in Lysine biosynthetic cluster
General Stress Response and Stationary Phase Response
Flagella protein
Organic hydroperoxide resistance related hypothetical protein
proteosome related
Hypothetical protein possible functionally linked with Alanyl-tRNA synthetase
Ribosome-related cluster
Cytochrome biogenesis
Chemotaxis, response regulators
Plant cell walls and outer surfaces
Pyruvate kinase associated cluster
Type III secretion system, extended
Degradation of Polyphenols
recA and regulatory cluster
Menaquinone--cytochrome C reductase
Membrane-bound hydrogenase
Carotenoid biosynthesis
Pigment biosynthesis
Recombination related cluster
Nucleotidyl-phosphate metabolic cluster
Hypothetical associated with RecF
heat shock, cell division, proteases, and a methyltransferase
Urate degradation
Chromosome Replication
CRISPRs and associated hypotheticals
Two related proteases
Hypothetical Related to Dihydroorate Dehydrogenase
Proteasome related clusters
RNA polymerase sigma-54 factor rpoN
Plastidial (cyanobacterial) electron transport system
Protein export
Tricarboxylate transporter
Secondary Metabolism
Phages, Prophages, Transposable elements
Ribosomal Protein L28P
Photosynthesis
Phosphorus Metabolism
Cell Division
Nitrogen Metabolism
Metabolism of Aromatic Compounds
Stress Response
RNA Metabolism
Cell Wall and Capsule
DNA Metabolism
Respiration
Carbohydrates
Miscellaneous
Amino Acids and Derivatives
Protein Metabolism
Cofactors, Vitamins, Prosthetic Groups, Pigments
ISL
DSP
DSS
MWC
CMS
SRB

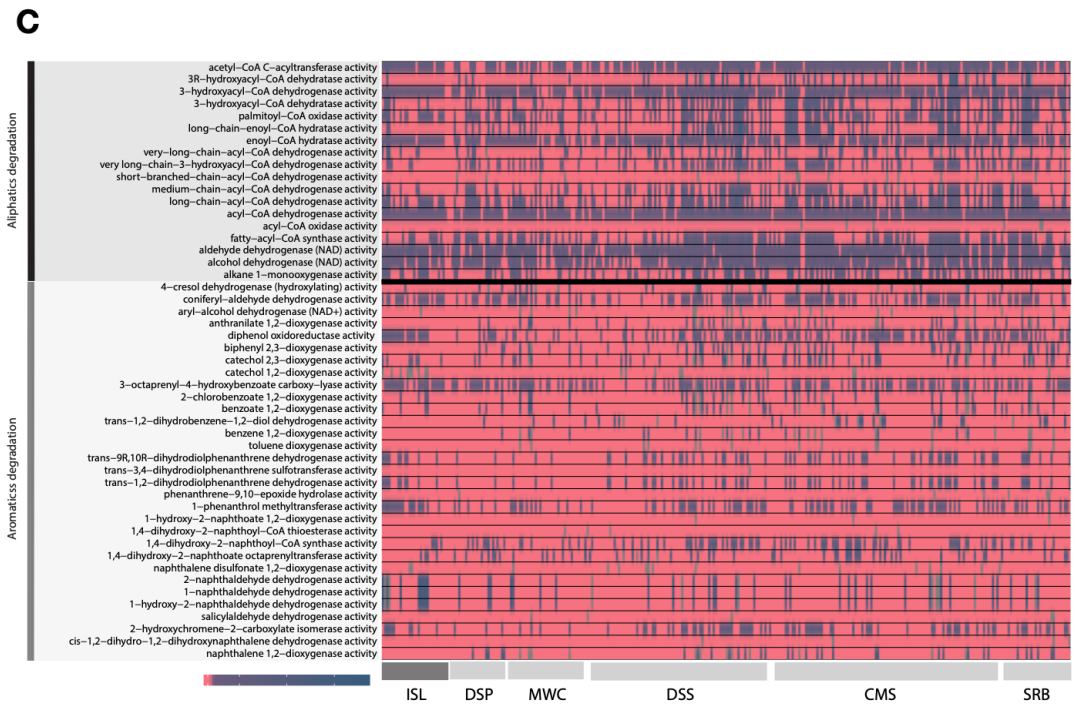


Figure 5-5: Heatmap showing SEED subsystem based clustering of the genes predicted in the metagenomes (A) and MAGs (B) clustered at 40% amino acid ID for the DWH impacted samples. C. Gene abundance for specific functions related to oil biodegradation detected in the MAGs. Note that no distinct clustering patterns are observed for the functional genes encoded by these MAGs. For simplicity the ecosystems have been broadly categorized as deep-sea sediment (DSS), deep-sea oil plume (DSP), coastal marine sediments (CMS), surface residue balls (SRB) and marine water column (MWC) for uncontaminated samples recovered from the Gulf of Mexico.

Despite recent efforts, many sites, pristine or contaminated, are under-sampled and additional environmental surveys are needed for a complete picture of oil-associated microbial diversity to emerge. The GROS project provides an easy integration of such data and capabilities to analyze new and unpublished data in the context of previously published data. This repository can be used to obtain a more holistic view of the microbial responses to oil, differences and similarities across habitats, and identify biomarkers that can be universally representative of the different phases of oil biodegradation and ecosystem recovery as exemplified above. The curated database may serve as a model approach for building similar resources for additional environmental processes and data of interest such as the waste/drinking water microbiome (Hull et al. 2019). GROS is publicly available through the link below and allows the user to download all MAGs and metadata mentioned above: http://microbial-genomes.org/projects/GROS.

5.5 Acknowledgements

This research was made possible by grants from The Gulf of Mexico Research Initiative (RFP V Grant No 321611-00 as well as grants to the C-IMAGE II, C-IMAGE III, and Deep-C consortia). Data are publicly available at http://microbial-genomes.org/. Associated NCBI accession numbers have also been provided where applicable. The authors would also like to thank Ramachandra Sivakumar form the Center for Spatial Planning Analytics and Visualization (CSPAV) at Georgia Institute of Technology for his valuable input with the incorporation of GIS layers in the webserver.

CONCLUSIONS AND FUTURE PERSPECTIVES

Hundreds of millions of liters of petroleum hydrocarbons enter the environment every year as a result of natural oil seeps or anthropogenic activities and accidents. Assessing the impacts of an oil spill on the environment entails an intrinsic understanding of both the movement and the physicochemical changes the oil undergoes after its release. Studies of the fate, transport, and potential for natural attenuation of oil and natural gas have been pressing topics in oil spill research and environmental assessment for decades. The physicochemical processes that oil undergoes once it is released into the environment range from dissolution, dispersion and evaporation to photodegradation (Murray, Boehm, and Prince 2020). Several large scale investigations following the *Amoco Cadiz* tanker spill (1978), the *Ixtoc I* well blowout (1979) and the *Exxon Valdez* spill (1989) led to the standardization of protocols for evaluating the impacts of shoreline oiling (Owens and Sergy 2003). The *Ixtoc I* well blowout in the Gulf of Mexico (Campeche Bay) was thought to have undergone rapid biodegradation due to the local microbial communities being adapted to natural seeps in the area. However, the emulsification of the oil reduced its bioavailability thus, preventing significant biodegradation (Atlas 1995). The Exxon Valdez spill was the first spill to use extensive bioremediation efforts as the main strategy for clean-up operations (Atlas and Hazen 2011). However, the lack of next generation sequencing data precluded further interpretations of the microbial community dynamics and nutrient limitations for microbial activity that controlled the ultimate fate of oil in the previously mentioned spills. The Deepwater Horizon oil spill was the first oil spill to be characterized with an unprecedented level of resolution using next-generation sequencing

technologies. “Omics” is a suite of extremely powerful tools in disentangling complex microbial community level interactions and structuring. For example, the tree of life before the omics era was significantly smaller.

While omics-based analyses have the potential to untangle complex microbial interactions in the environment, integration of *in-situ* transport processes as well as flux measurements could greatly augment the omics-based modelling of microbial community response toward better forecasting of the fate of spilled crude-oil. This integrated approach currently remains a missing link in advancing our understanding of the microbial controls of oil degradation in the environment. Development of computational models is increasingly important in our understanding of the fate of spilled oil, because they can fill in gaps in time and space for which data are unavailable or inaccessible. However, these tools are severely limited based on the quality of the input data and ultimately depend on incorporating the field and laboratory-based measurements. Although forecasting precise biodegradation rates of spilled oil in any given place and time is extremely difficult, good experimental data generated from robust laboratory incubations, using realistic *in-situ* conditions like temperature, flow processes and nutrient levels, can be used to incorporate biodegradation processes into conceptual models, resulting in improved prediction of oil fate. Further, an easily accessible database documenting different oil spill incidents and their biogeochemical data could greatly aid in how we plan emergency spill responses. Having this data readily and easily available would also help in generating more robust models to track the fate of oil as well as providing robust baseline data. GoM’s ERMA (https://response.restoration.noaa.gov/gulf-mexico-erma) and GROS (http://microbial-genomes.org/projects/GROS) are examples of such tools to link microbiological data to

real-time environmental data toward better planning ecological recovery and restoration efforts. The database curated as part of this thesis should facilitate future studies to further understand the interactions among microbial community members and their chemical environment that ultimately control the fate of oil spills. Furthermore, our large-scale analysis of next generations sequencing data generated reliable biomarkers to screen for oil degradation potential in marine ecosystems which are essential in determining if an ecosystem is more "primed" for oil biodegradation.

The Deepwater Horizon oil spill is also one of the best examples of how the *in-situ* flow dynamics impacted the microbial community structuring at all locations impacted by the spill. The addition of the chemical dispersant Corexit 9500 further affected the solubility and transport of oil. The discharge, which originated at a depth of 1500 m below the ocean surface, underwent kinetic fractionation which significantly altered the composition of the dispersed oil and in turn, the microbial community at different locations of the oil-plume (Valentine et al. 2012; King et al. 2015). A coupled physical-metabolic model developed previously revealed that current oscillation and mixing processes were instrumental in the distribution of hydrocarbons and their associated bacterial community in the nGoM (Valentine et al. 2012). An "autoinoculation" effect was induced through priming of the indigenous microbiota residing in water masses that recirculated around the wellhead. These transport-microbial coupled modelling-based approaches could be used not only in tracking the speed/efficacy of oil degradation but also in determining the controlling factors of degradation at a given location. The deep-sea conditions (where the spill originated) were markedly different in terms of physicochemical conditions compared to the coastal sediments (where the discharge finally culminated).

This thesis primarily focused on coastal beach sediments because they were less studied relative to the deep-sea and surface ecosystems. Microbial and geochemical hydrocarbon modifications during the degradation process resulted in two-fold higher complexity of the buried oil that reached the shores when compared to the source well. The impact of the oil spill on coastal ecosystems remains comparatively less understood due to the stochasticity and complexity of ecosystem processes and lack of appropriate model microorganisms. This thesis explored the effect of transport processes in beach sands where transport of oxygen and nutrients is expected to control the rates of hydrocarbon biodegradation. The integration of transport processes in designing and operating advective chambers to study microbial oil degradation in the field led to remarkably similar microbial communities to those observed *in-situ*. Using genome-resolved metagenomic approaches, the key hydrocarbon degrading and nitrogen-fixing microorganism in these laboratory incubations, which made up ~30% of the total microbial community, was isolated and characterized. This organism, provisionally named "*Candidatus* Macondimonas diazotrophica", represents a previously overlooked family of hydrocarbon degraders that are major responders to oil spills in coastal environments worldwide. *Ca*. Macondimonas is a promising model organism for studying ecophysiological responses to oil spills.

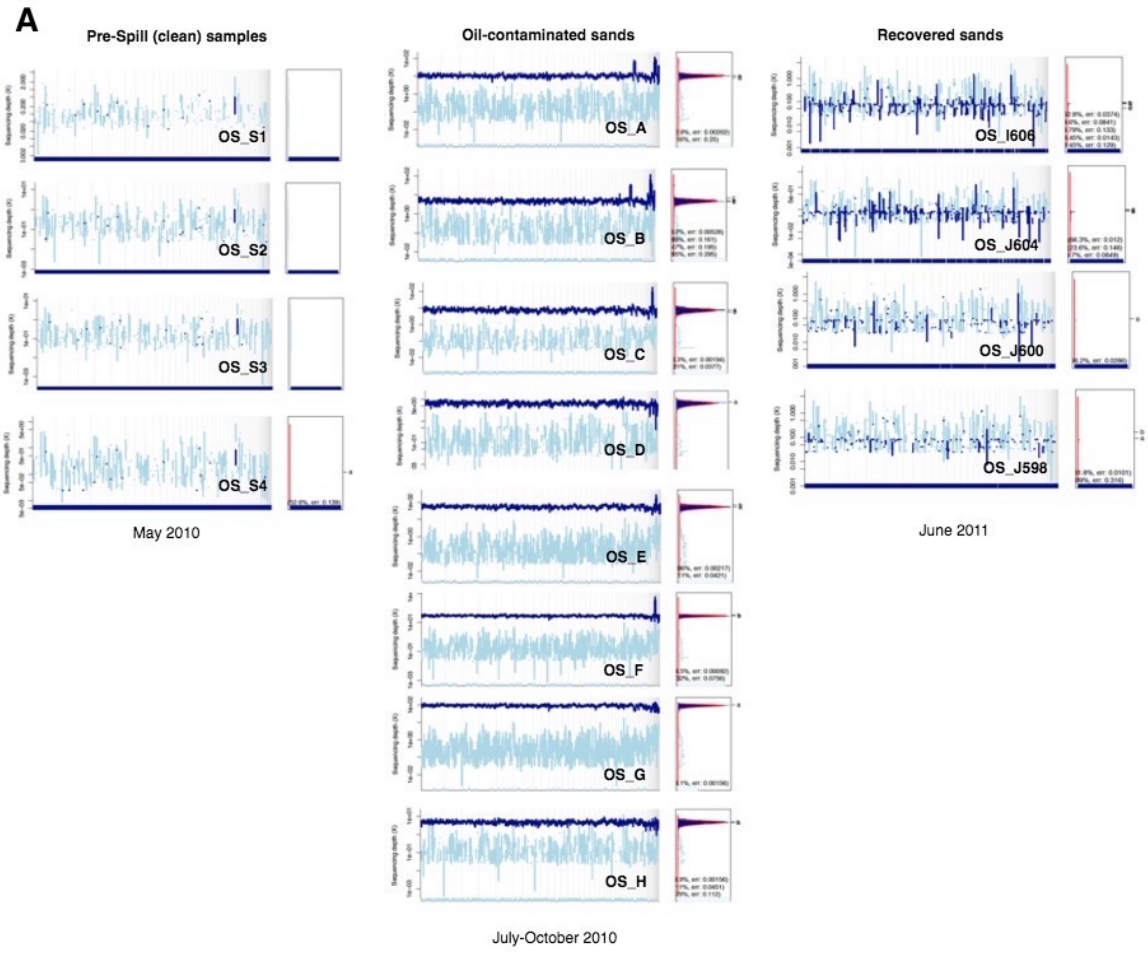
The thesis also helped shed light on which species or genes are crucial at particular time points and their significance in degrading buried crude oil. Furthermore, it offered critical insight into the microbial community dynamics and its impact of oil degradation when subject to periods of anoxia. This could form the basis for developing models using a systems approach that incorporates microbiological, genomics, biogeochemical and transport data to predict decomposition rates of buried oil in sub-, inter and supratidal beach

sands. For instance, observations based on population sizes and gene abundances, measured at sequential time points, can be integrated into dynamic mathematical models, using established approaches such as the modified Lotka-Volterra (LV) models along with methods of slope estimation, decoupling and linear regression.

Currently, experiments employing laboratory systems that simulate the diurnal tidal cycles are underway in the Konstantinidis Lab to explore the microbial community dynamics in temporally saturated coastal sediments. These integrated approaches could be a powerful tool in disentangling complex ecosystems interactions as well as provide a multiphasic, systems-level approach to not only oil-impacted ecosystems but also other environmentally relevant processes; for instance, to monitor the fate and transport of environmentally relevant pollutants.

APPENDIX A. SUPPLEMENTAL MATERIAL FOR CHAPTER 2

A.1 Supplemental figures and tables



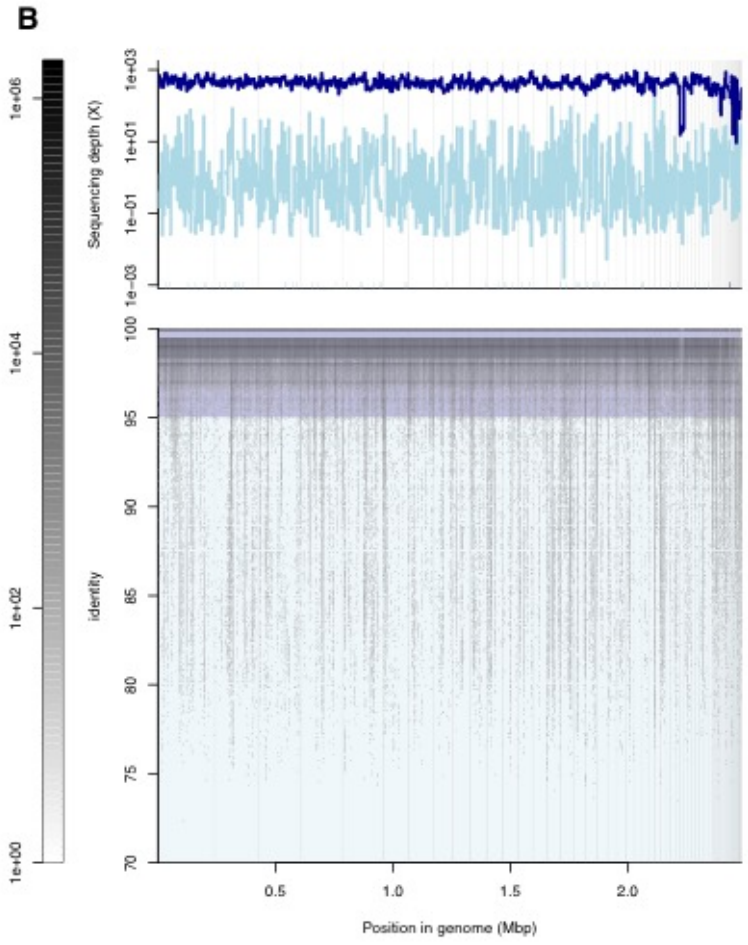


Figure A 1: A. Coverage of the MAG-01 by each metagenome. Average coverage, representing relative abundance of MAG-01 sequence (x-axis) by the reads of the metagenomic datasets described in Rodriguez et al. 2015 (y-axis). The dark blue histogram represents reads matching the reference genome at ≥ 80 bp in length and ≥ 95% nucleotide identity; light blue represents reads matching at < 95% identity. Note that the genome has low abundance or is absent in pre-spill and recovered metagenomes. B: Read recruitment plots for MAG-01 by a metagenome obtained from coastal Alabama beach sands (Dauphin Island) (MG-RAST Accession ID: mgp8963). Top panel: dark blue histogram represents

reads matching the reference genome at ≥ 80 bp in length and ≥ 95% nucleotide identity; light blue represents reads matching at < 95% identity. Bottom panel shows the read recruitment plots where the metagenomic reads matched to MAG-01 along with the % nucleotide identity of the match. The evenness of the coverage of the genome on the metagenomic datasets shows a sequence discrete population. Note that the graphs on panel A show only the average coverage of the reference MAG-01 sequence by reads (top of the graph in panel B), not the actual plot and where reads map (bottom of the graph in panel B).

Table A 1: Datasets with detectable abundances of KTK-01 16S rRNA gene sequence and its global distribution.

Country	Environment	ID	Accession	Reference
India	Hydrocarbon contaminated soil	99%	JN217180.1	
China	PAHs contaminated soil	98%	MF314818.1	
Romania	Oil-polluted soil	97%	DQ378238.1	
China	Soil contaminated with HCs	97%	KF912995.1	
China	Petroleum-contaminated soil	97%	JN038293.1	
France	Oil contaminated sediments	97%	FM242404.1	(Paisse et al 2010)
United States	Asphalt seeps	98%	DQ001646.1	
United States	Heavy oil seeps of the Rancho La Brea tar pits	98%	EF157259.1	(Kim and Crowley 2007)
China	Hydrocarbon contaminated saline-alkali soil	97%	HQ697765.1	
Germany	Metalworking fluid (petroleum distillates) contaminated sediment	97%	HE576011.1	Lodders & Kämpfer, 2012
Germany	Mineral oil contaminated	97%	HE576391.1	(Lodders and Kampfer 2012)
Antarctica	Ornithogenic soil affected by oil spills	97%	FJ380163.1	(Aislabie et al 2008)
Bulgaria	Non-ferrous metal smelting factory KCM	97%	AJ581598.1	
France	Pilot-scale bioremediation process of a hydrocarbon-contaminated soil	97%	FM209143.1	
Tunisia	Constructed wetland down stream of active sludge	97%	KC432416.1	(Bouali et al 2014)
Spain	Hydrocarbon contaminated sand after the Prestige oil spill	97%	EU375160.1	(Alonso-Gutiérrez et al 2009)

Table A 2: Publicly available whole genome sequences of oil-degrading isolates used for comparison with isolate KTK-01.

Order	Genus	Sp./Strain	Conditions	Alkane degrading genes	Nitrogenase genes	Ref.
Oceanospirillales	*Alcanivorax*	PN3	Anaerobic	✓	×	Overholt et al., 2013
Oceanospirillales	*Alcanivorax*	P2S70	Aerobic	✓	×	Overholt et al., 2013
Alteromonadales	*Marinobacter*	EN3	Anaerobic	✓	×	Overholt et al., 2013
Alteromonadales	*Marinobacter*	EVN1	Anaerobic	✓	×	Overholt et al., 2013
Alteromonadales	*Marinobacter*	C1S70	Aerobic	✓	×	Overholt et al., 2013
Alteromonadales	*Acinetobacter*	COS3	Aerobic	✓	×	Overholt et al., 2013
Oceanospirillales	*Halomonas*	PBN3	Anaerobic	✓	×	Overholt et al., 2013
Rhodobacter	*Labrenzia*	C1B10	Aerobic	✓	×	Overholt et al., 2013
Rhodobacter	*Labrenzia*	C1B70	Aerobic	✓	×	Overholt et al., 2013
Oceanospirillales	*Alcanivorax*	SK2	Aerobic	✓	×	Schneiker et al, NatureBT, 2006
Alteromonadales	*Marinobacter*	*hydrocarbonoclasticus*	Aerobic	✓	×	Grimaud et al, 2012
Alteromonadales	*Marinobacter*	*aquaeolei VT8*	Aerobic	✓	×	
Alteromonadales	*Marinobacter*	*hydrocarbonoclasticus*	Aerobic	✓	×	Present study
Alteromonadales	*Microbulbifer*		Aerobic	✓	×	Present study
Gammaproteobacteria		*KTK-01*	Aerobic	✓	✓	Present study
Rhodobacter	*Hyphomonas*		Aerobic	×	×	Present study
Oceanospirillales	*Kangiella*		Aerobic	✓	×	Present study

Table A 3: Putative genes identified in the genome encoding major metabolic pathways. For selected genes, their amino acid identity (4th column) and closest verified homolog (5th and 6th column) are also shown.

Gene		AA length	%ID	Closest hit	Accession
Nitrogen fixation					
nifH	Nitrogenase (molybdenum-iron) reductase	297	89.7	*Azorhizobium caulinodans*	P26251 (SwissProt)
nifD	Nitrogenase molybdenum-iron protein alpha chain	485	84	*Azorhizobium caulinodans*	WP_012169570.1
nifK	Nitrogenase molybdenum-iron protein beta chain	518	77.4	*Bradyrhizobium diazoefficiens*	P20621 (SwissProt)
nifX	Nitrogenase FeMo-cofactor carrier protein	164	80	*Acidithiobacillus ferrivorans*	WP_065413794.1
nifN	Nitrogenase FeMo-cofactor scaffold and assembly protein	453	61.1	*Bradyrhizobium diazoefficiens*	P26507
nifE	Nitrogenase FeMo-cofactor scaffold and assembly protein	482	83	*Methylocystis spp.*	WP_014892105.1
nifQ	Nitrogenase FeMo-cofactor synthesis molybdenum delivery protein	209	48	*Cupriavidus basilensis*	WP_059410667.1
	4Fe-4S ferredoxin, nitrogenase-associated				
Aliphatics degradation and beta oxidation					
*alkB**	Alkane-1 monooxygenase	410	46	*Alcanivorax sp.*	WP_035460937.1
alkB1	Alkane-1 monooxygenase	392	45	*Marinobacter sp. ES-1*	WP_022988941.1
alkB2	Alkane-1 monooxygenase		31.3	*Alcanivorax borkumensis*	
*adh**	Alcohol dehydrogenase	333	70	*Gammaproteobacterium SG8_31*	KPK60815.1
	Aldehyde dehydrogenase	488	45.4	*Pseudomonas aeruginosa*	Q9I6C8
	Alkanal monooxygenase				

Table A 3 continued

	Long-chain fatty-acid CoA ligase				
	Medium-chain fatty-acid CoA ligase				
	Acyl-CoA dehydrogenase				
	Medium-chain specific acyl-CoA dehydrogenase				
	Acyl-CoA dehydrogenase, short-chain specific				
	Enoyl-CoA hydratase				
	3-hydroxyacyl-CoA dehydrogenase				
	3-ketoacyl-CoA thiolase				
	Acetyl-CoA synthetase				
	Acetyl-CoA acetyltransferase				
	Acyl-[acyl-carrier-protein] synthetase				
	** Multiple copies found*				
Methanotrophy					
pmoC	Particulate methane monooxygenase C-subunit	225	41	*Methylococcus capsulatus*	AAB49820.1
pmoA	Particulate methane monooxygenase A-subunit	262	43.04	*Methylococcus capsulatus*	WP_010961050.1
pmoB	Particulate methane monooxygenase B-subunit	438	40	*Methylococcus capsulatus*	WP_010961049.1
	Methenyltetrahydrofolate cyclohydrolase				
glyA	Serine hydroxymethyltransferase				
	5,10-methylenetetrahydrofolate reductase				
Biosurfactant production					
ompA	OmpA-like protein	544	30	*Gammaproteobacterium*	PKM18649.1

Table A 3 continued

oprF	Outer membrane porin F	361	40.49	*Pseudomonas aeruginosa*	WP_034002471.1
SOS response and Osmoregulation					
DNArepair					
uvrC	UvrABC system protein C				
uvrA	UvrABC system protein A				
uvrB	UvrABC system protein B				
recA	Protein RecA				
recQ	ATP-dependent DNA helicase				
recN	DNA repair protein				
Osmoregulation					
osmY	Osmotically-inducible protein Y				
mnhC	Na(+)/H(+) antiporter subunit C				
mnhB	Na(+)/H(+) antiporter subunit B				
mnhF	Na(+)/H(+) antiporter subunit F				
mnhE	Na(+)/H(+) antiporter subunit E				
Urea Metabolism					
ureA	Urease subunit gamma				
ureE	Urease accessory protein UreE				
ureC	Urease subunit alpha				

Table A 3 continued

ureB	Urease subunit beta
TCA cycle	
gltA	Citrate synthase
acn	Aconitate hydratase A
aceA	**Isocitrate lyase (Glyoxylate shunt)**
aceB	**Malate synthase A (Glyoxylate shunt)**
icd	Isocitrate dehydrogenase [NADP]
sucA	2-oxoglutarate dehydrogenase E1 component
sucB	Dihydrolipoyllysine-residue succinyltransferase component
sucC	Succinate--CoA ligase [ADP-forming] subunit beta
sucD	Succinate--CoA ligase [ADP-forming] subunit alpha
sdhA	Succinate dehydrogenase flavoprotein subunit
sdhC	Succinate dehydrogenase cytochrome b556 subunit
sdhD	Succinate dehydrogenase hydrophobic membrane anchor subunit
sdhB	Succinate dehydrogenase iron-sulfur subunit
mdh	Malate dehydrogenase
fumA	Fumarate hydratase class I, aerobic
Secretion and biofilm formation	
Type II secretion systems	

Table A 3 continued

gspC	Type II secretion system protein C
gspD	Type II secretion system protein D
gspE	Type II secretion system protein E
gspF	Type II secretion system protein F
gspG	Type II secretion system protein G
gspH	Type II secretion system protein H
gspI	Type II secretion system protein I
gspJ	Type II secretion system protein J
gspK	Type II secretion system protein K
gspL	Type II secretion system protein L
gspM	Type II secretion system protein M
Sec translocon	
yajC	Membrane protein
secD	Protein translocase subunit
secF	Protein translocase subunit
secA	Protein translocase subunit
secB	Protein-export protein
secG	Protein-export membrane protein
Twin arginine traslocation system	
tatA	Sec-independent protein translocase protein
tatB	Sec-independent protein translocase protein
tatC	Sec-independent protein translocase protein

Table A 3 continued

Type IV pilus	
pilF	Type IV pilus biogenesis protein PilF
pilZ	Type IV pilus biogenesis protein PilZ
pilJ	type IV pilus biogenesis protein PilJ
pilM	Type IV pilus biogenesis protein PilM
pilN	Type IV pilus biogenesis protein PilN
pilO	Type IV pilus biogenesis protein PilO
pilP	Type IV pilus biogenesis protein PilP
pilQ	Type IV pilus biogenesis protein PilQ
pilE	Type IV pilus biogenesis protein PilE
pilB	Type IV fimbrial assembly protein PilB
pilT	Twitching mobility protein

APPENDIX B. SUPPLEMENTAL MATERIAL FOR CHAPTER 3

B.1 Supplementary Figures and Tables

Table B 1: 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

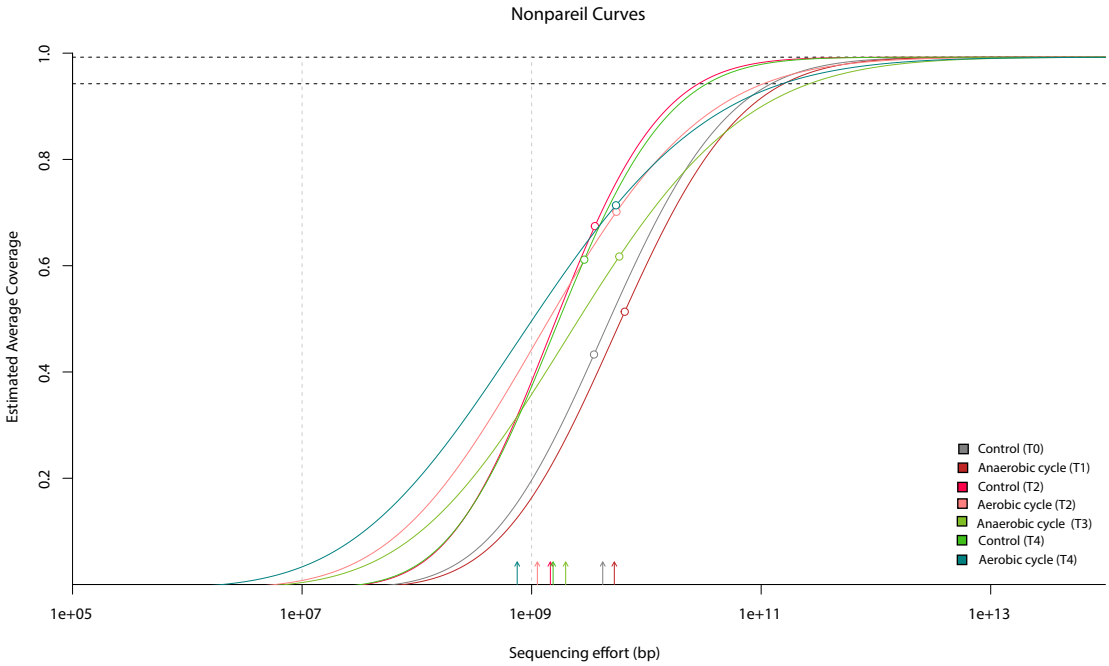


Figure B 1: Nonpareil diversity and sequencing coverage estimates of the microbial communities. Empty circles represent the estimated average coverage of the datasets obtained and projections based on model fitting to reach 95% and 99% coverage are indicated (horizontal dashed lines. The arrows at the bottom represent sequencing effort required to achieve 50% coverage.

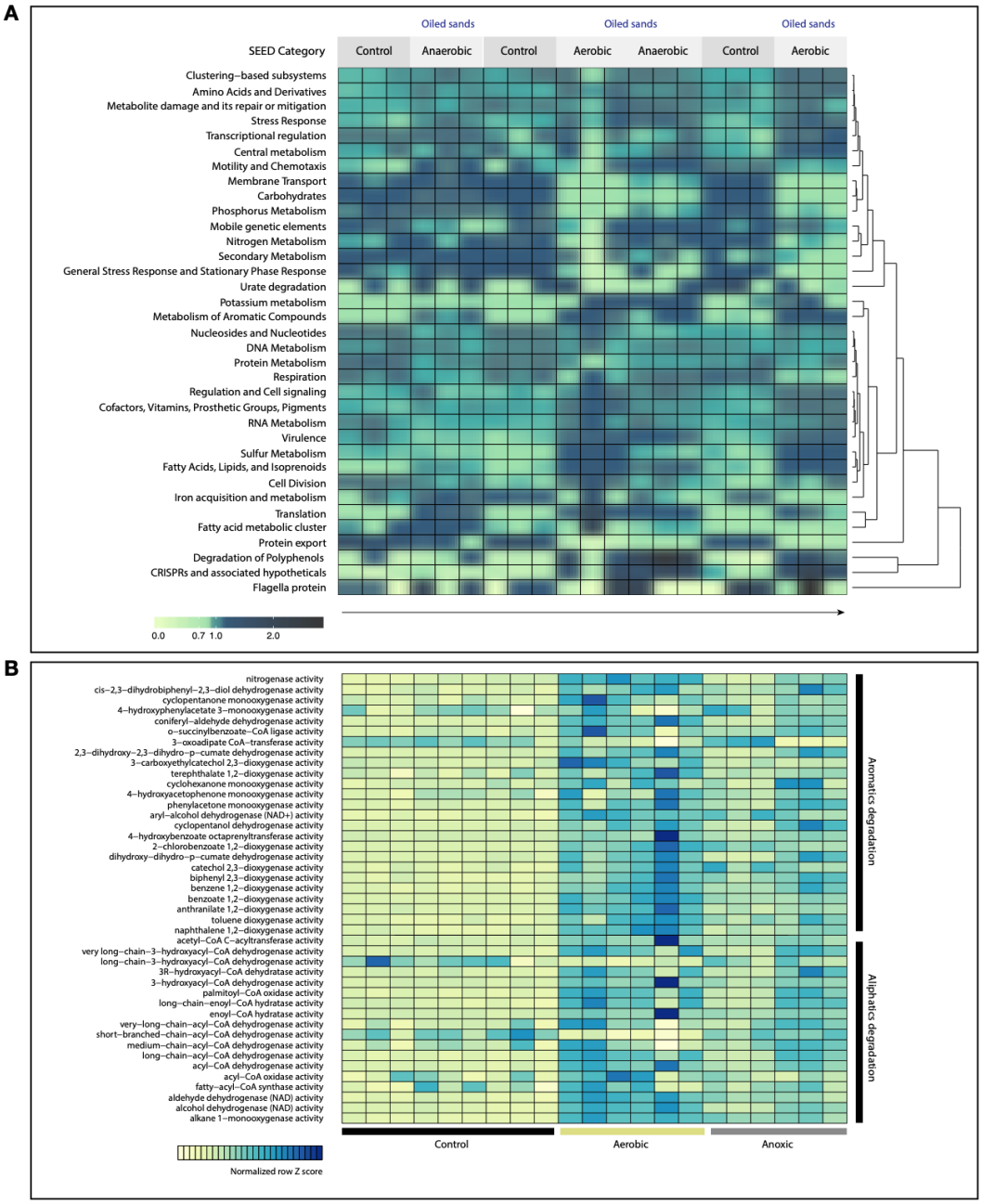


Figure B 2: Read-level functional gene shifts during the mesocosm incubation. Heatmaps showing the relative abundance of major metabolic pathways (Level 1 of the SEED subsystems; Panel A) and individual gene functions related to oil biodegradation (Panel B).

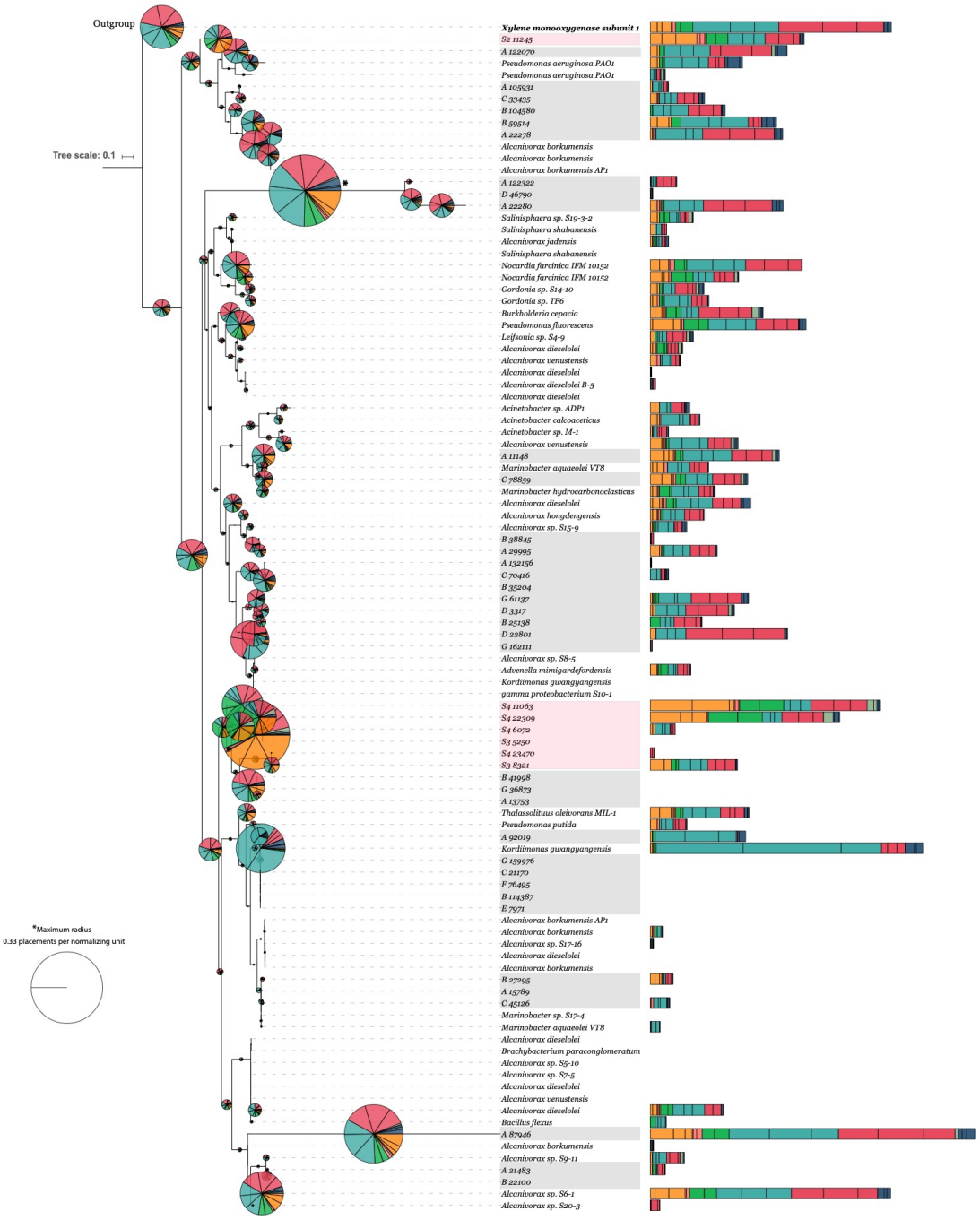


Figure B 3: Phylogenetic diversity of *AlkB*-containing metagenomic reads. The radii of the pie charts and bars are proportional to the normalized number of reads assigned to ta particular clade. Clades highlighted in gray indicate putative sequences recovered from

the field datasets. Clades highlighted in red indicate putative sequences recovered uncontaminated field sites. Outgroup: xylene monooxygenase *XylM*.

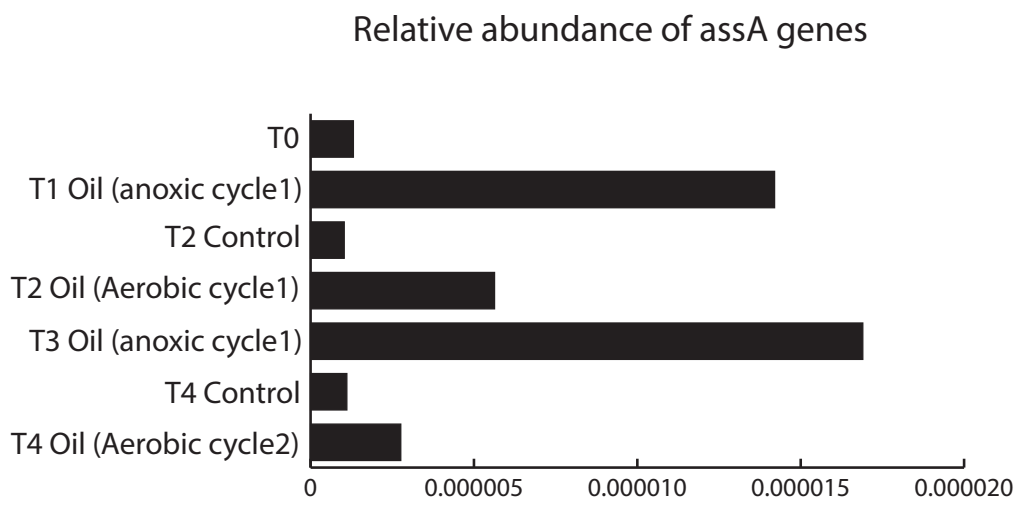


Figure B 4: Alkyl/methylsuccinate synthase (*assA*) gene relative abundance in the metagenomic datasets. Gene abundance normalized by metagenome library size.

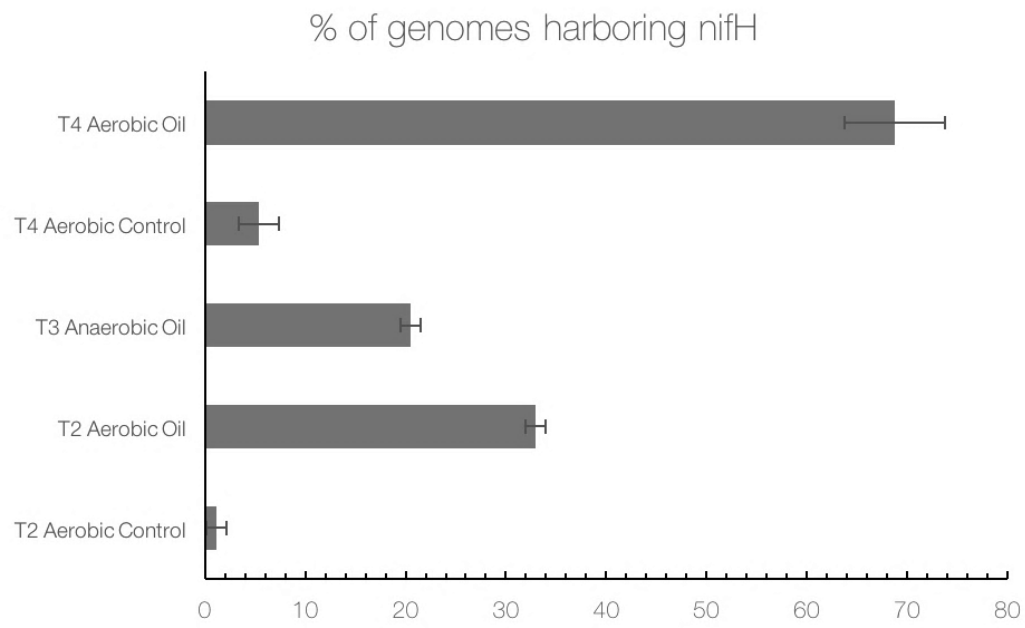
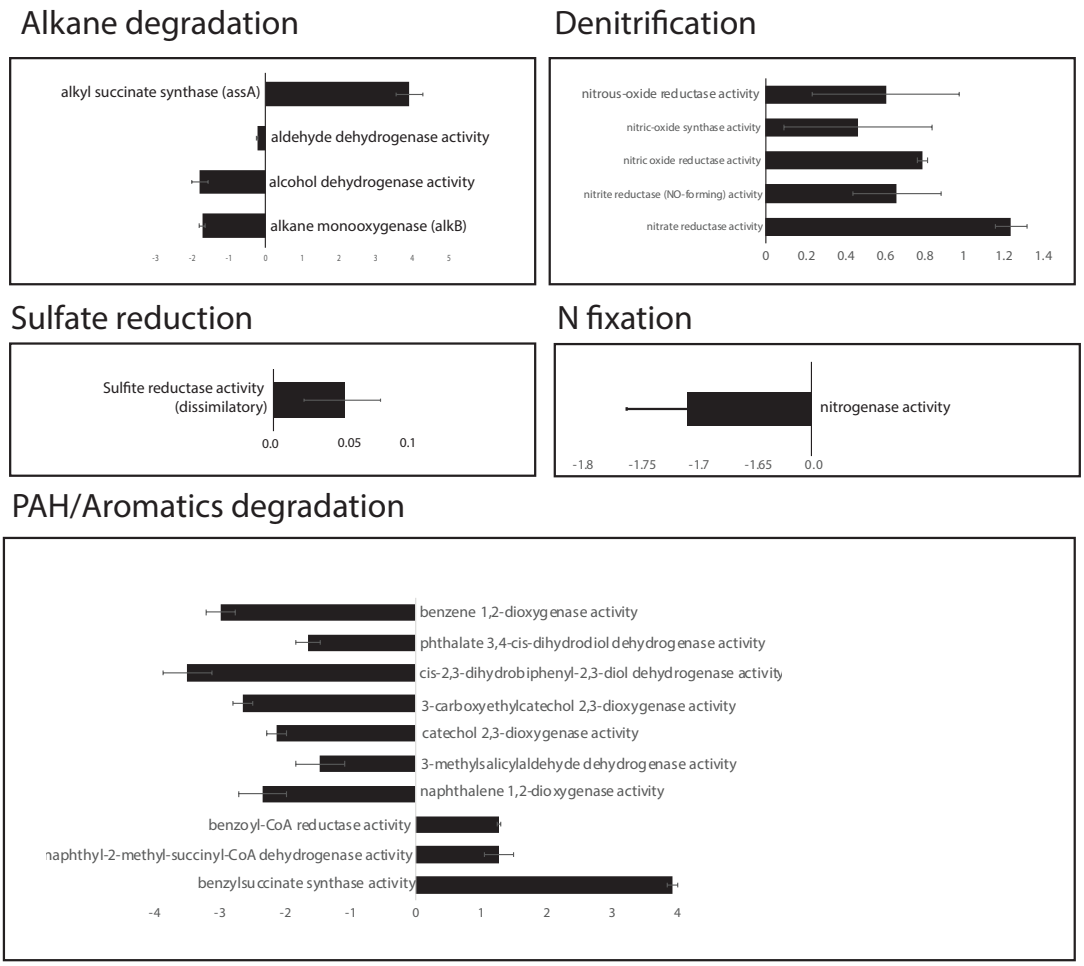


Figure B 5: Fraction as genomes harboring the nitrogenase (*nifH*) gene through the time course of the study. Normalized gene abundance was calculated as genome equivalents (i.e. normalized by abundance of the single copy *rpoB* gene). Mean of the 3 biological replicates and their SD are shown here.


Alkane degradation
alkyl succinate synthase (assA)
aldehyde dehydrogenase activity
alcohol dehydrogenase activity
alkane monooxygenase (alkB)
-3 -2 -1 0 1 2 3 4 5
Denitrification
nitrous-oxide reductase activity
nitric-oxide synthase activity
nitric oxide reductase activity
nitrite reductase (NO-forming) activity
nitrate reductase activity
0 0.2 0.4 0.6 0.8 1 1.2 1.4
Sulfate reduction
Sulfite reductase activity
(dissimilatory)
0.0 0.05 0.1
N fixation
nitrogenase activity
-1.8 -1.75 -1.7 -1.65 0.0
PAH/Aromatics degradation
benzene 1,2-dioxygenase activity
phthalate 3,4-cis-dihydrodiol dehydrogenase activity
cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase activity
3-carboxyethylcatechol 2,3-dioxygenase activity
catechol 2,3-dioxygenase activity
3-methylsalicylaldehyde dehydrogenase activity
naphthalene 1,2-dioxygenase activity
benzoyl-CoA reductase activity
naphthyl-2-methyl-succinyl-CoA dehydrogenase activity
benzylsuccinate synthase activity
-4 -3 -2 -1 0 1 2 3 4
More in aerobic
More in anoxic
Log 2-Fold Change

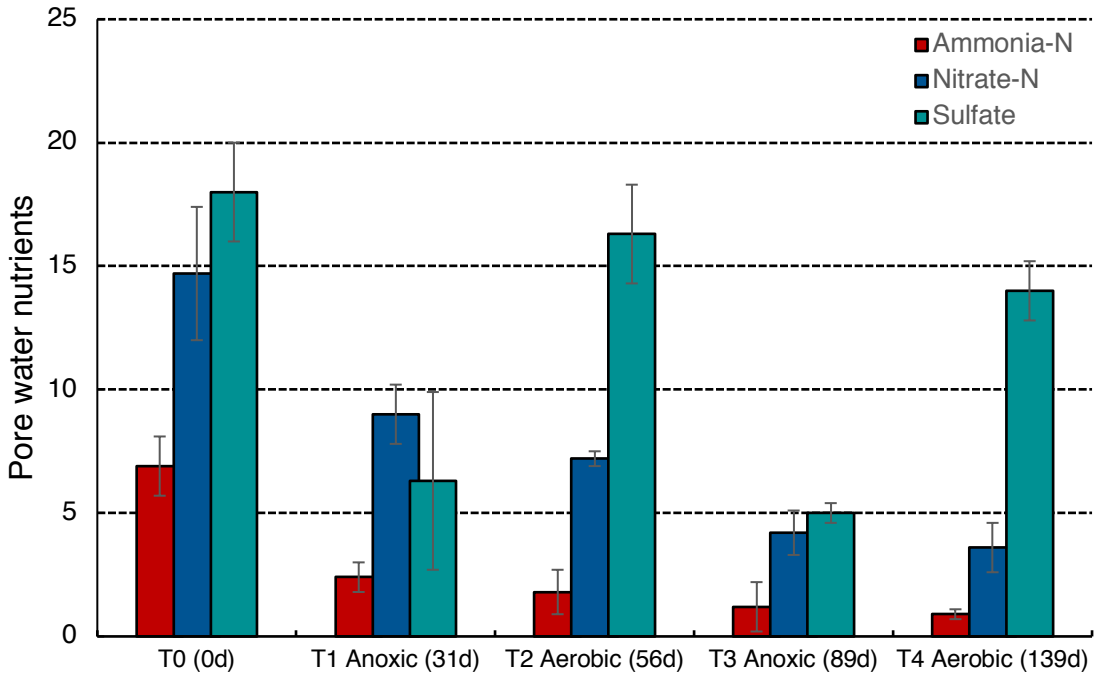


Figure B 6: Upper panel: Log 2-fold changes in the expression of the associated pathway genes shown in Figure 3-5. Mean and SD of the three biological replicates are reported here. **Lower panel: Measured pore-water nutrient concentration in the oiled chambers.** Mean and SD of the three biological replicates 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.

Table B 2: 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 B 3: 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*	48.4431190
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	*Woeseia oceani NZ CP016268*	48.3105055
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*	46.1579590
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*	75.5178380
T2 Oil Bins 034	*Thiohalobacter thiocyanaticus NZ AP018052*	45.1892200
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	*Candidatus Campbellbacteria bacterium GW2011*	42.9542287
T3 Oil Bins 003	*Thiohalobacter thiocyanaticus NZ AP018052*	46.2457598
T3 Oil Bins 004	*Parvibaculum lavamentivorans DS 1 NC 009719*	50.3055530
T3 Oil Bins 005	*Alcanivorax sp NBRC 101098 NZ AP014613*	79.6975540
T3 Oil Bins 006	*Alcanivorax sp NBRC 101098 NZ AP014613*	76.4814925
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	*Candidatus Campbellbacteria bacterium GW2011*	41.2911452
T3 Oil Bins 016	*Thiohalobacter thiocyanaticus NZ AP018052*	48.8381799
T3 Oil Bins 018	*Desulfococcus oleovorans Hxd3 NC 009943*	47.7848973

Table B 3 continued

T3 Oil Bins 019	*Muricauda lutaonensis NZ CP011071*	62.6401100
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*	55.7286390
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*	48.2386030
T4 Oil Bins 026	*Candidatus Hodgkinia cicadicola CP024746*	41.8003557
T4C Bins 001	*Zobellia galactanivorans NC 015844*	57.7536190
T4C Bins 002	*Thiohalobacter thiocyanaticus NZ AP018052*	48.3222583

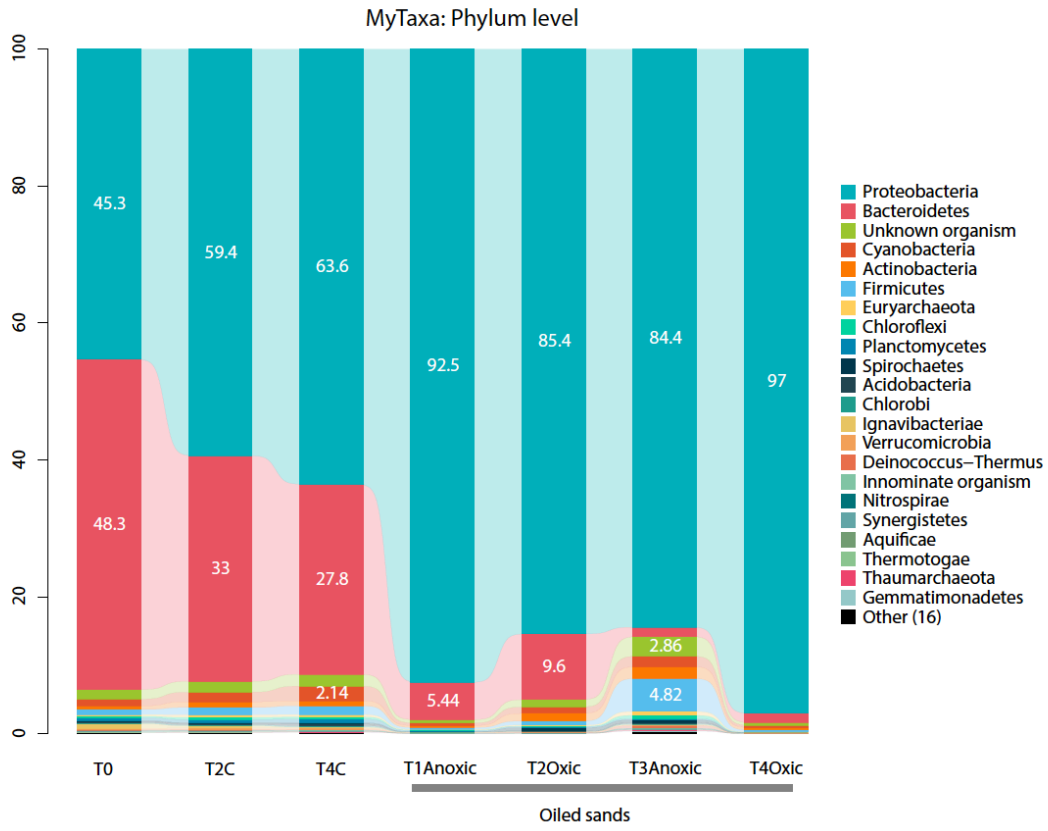


Figure B 7: MyTaxa classification of the co-assembled reads from each sampling time point.

APPENDIX C. SUPPLEMENTAL MATERIAL FOR CHAPTER 4

C.1 Supplementary Figures

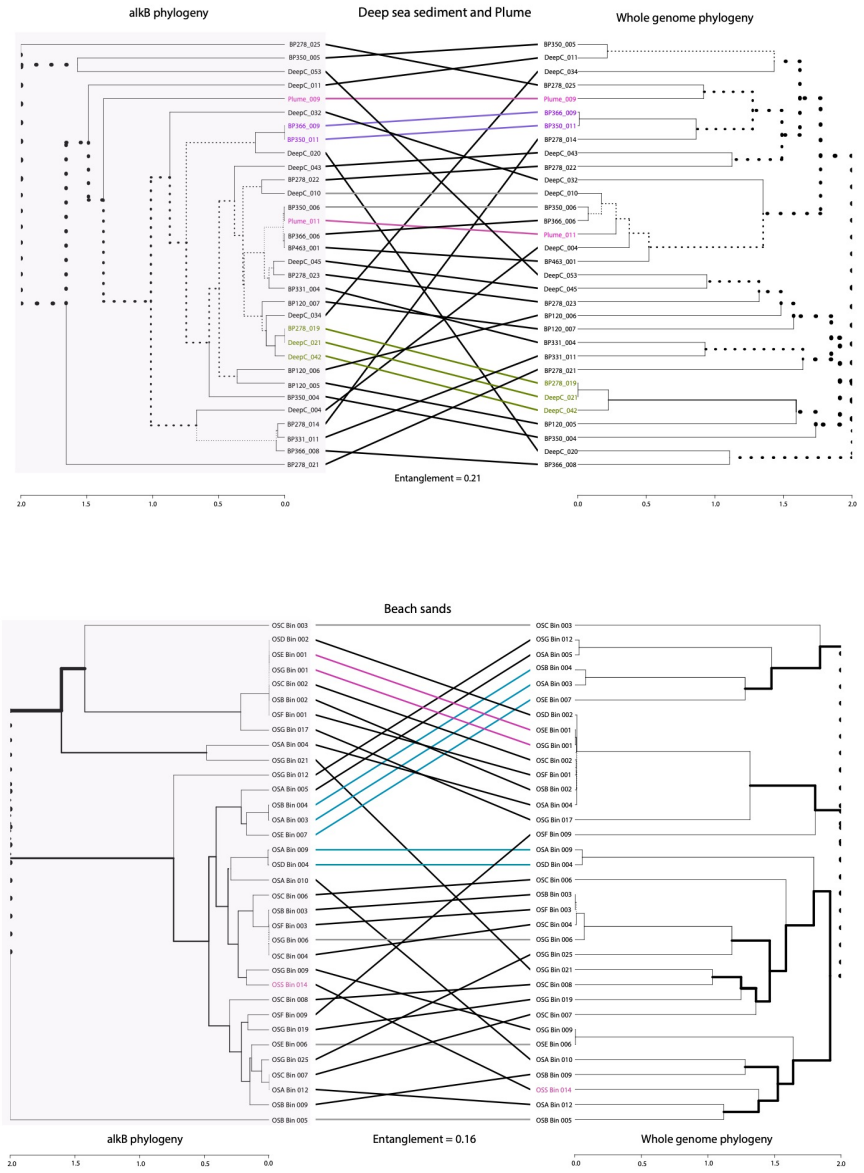


Figure C 1: Tanglegrams of the *alkB* gene (left) and whole genome (right) phylogeny of their corresponding MAGs. Top, beach sands; Bottom, deep sediments and water plume. No HGT were detected between beach sands and deep sediments and water plume.

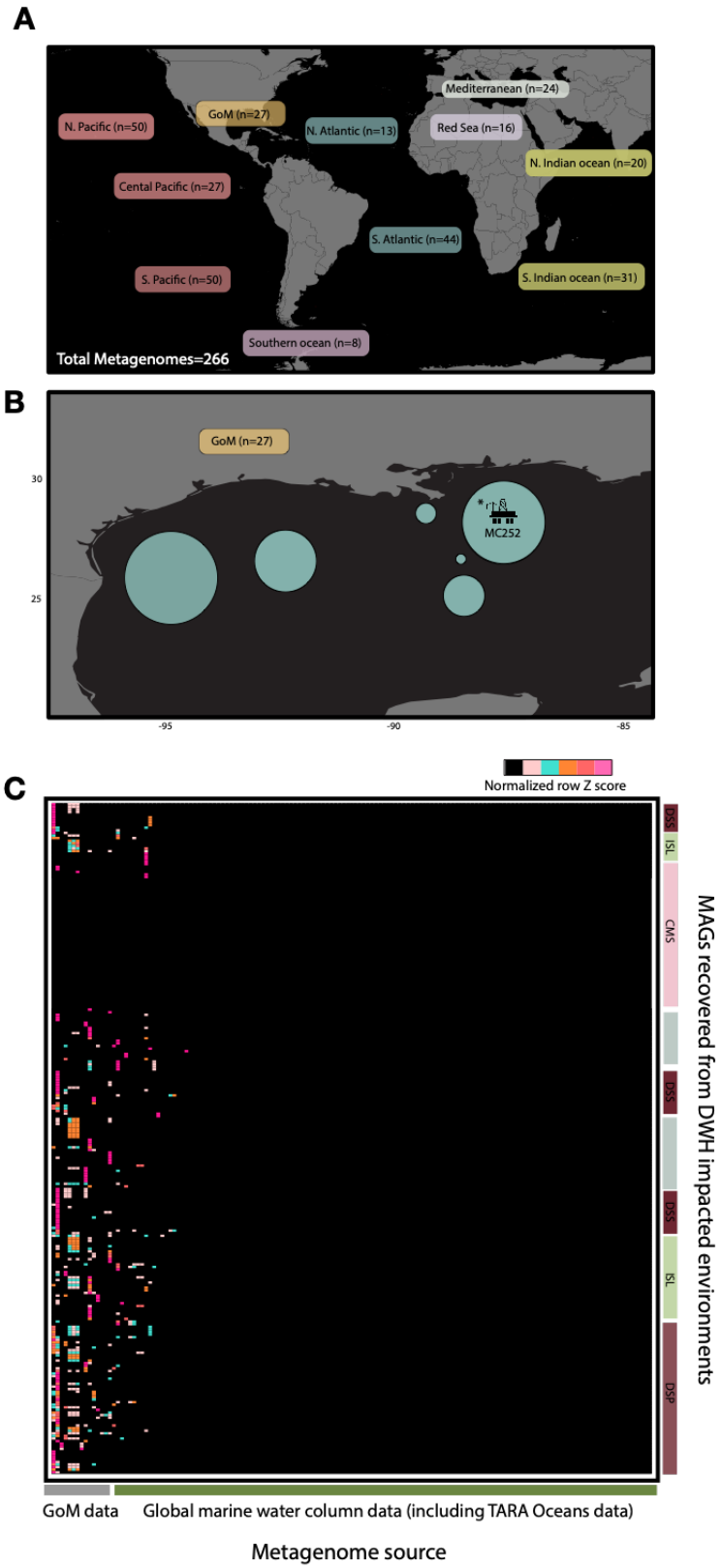

A
Mediterranean (n=24)
GoM (n=27)
N. Pacific (n=50)
N. Atlantic (n=13)
Red Sea (n=16)
N. Indian ocean (n=20)
Cental Pacific (n=27)
S. Atlantic (n=44)
S. Pacific (n=50)
S. Indian ocean (n=31)
Southern ocean (n=8)
Total Metagenomes=266
B
GoM (n=27)
MC252
30
25
-95
-90
-85
C
Normalized row Z score
DSS
ISL
CMS
DSS
DSS
ISL
DSP
MAGs recovered from DWH impacted environments
GoM data
Global marine water column data (including TARA Oceans data)
Metagenome source

Figure C 2: Geospatial distribution of the DWH MAGs in global ocean metagenomes recovered. A. Source of metagenome data (n=266) used in the study and the number of samples associated with each location. B. Similar to panel B but showing only the sampling sites associated with the GoM. C. Heatmap showing normalized abundance of the DWH impacted MAGs across the 266 metagenomes. For reference the 1st two columns show 2 of the metagenomes from the deep sea oil plume. For simplicity the ecosystems have been broadly categorized as deep-sea sediment (DSS), deep-sea oil plume (DSP), coastal marine sediments (CMS) and marine water column (MWC) for uncontaminated samples recovered from the Gulf of Mexico.

REFERENCES

Aller, Robert C. 1994. 'Bioturbation and remineralization of sedimentary organic matter: effects of redox oscillation', *Chemical Geology*, 114: 331-45.

Anantharaman, Karthik, Christopher T. Brown, Laura A. Hug, Itai Sharon, Cindy J. Castelle, Alexander J. Probst, Brian C. Thomas, Andrea Singh, Michael J. Wilkins, Ulas Karaoz, Eoin L. Brodie, Kenneth H. Williams, Susan S. Hubbard, and Jillian F. Banfield. 2016. 'Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system', *Nature Communications*, 7: 13219.

Atlas, R. M., and T. C. Hazen. 2011. 'Oil biodegradation and bioremediation: a tale of the two worst spills in U.S. history', *Environ Sci Technol*, 45: 6709-15.

Atlas, Ronald M. 1995. 'Petroleum biodegradation and oil spill bioremediation', *Marine Pollution Bulletin*, 31: 178-82.

Atlas, Ronald M., and Richard Bartha. 1972. 'Degradation and mineralization of petroleum in sea water: Limitation by nitrogen and phosphorous', *Biotechnology and Bioengineering*, 14: 309-18.

Aziz, Ramy K., Daniela Bartels, Aaron A. Best, Matthew DeJongh, Terrence Disz, Robert A. Edwards, Kevin Formsma, Svetlana Gerdes, Elizabeth M. Glass, Michael Kubal, Folker Meyer, Gary J. Olsen, Robert Olson, Andrei L. Osterman, Ross A. Overbeek, Leslie K. McNeil, Daniel Paarmann, Tobias Paczian, Bruce Parrello, Gordon D. Pusch, Claudia Reich, Rick Stevens, Olga Vassieva, Veronika Vonstein, Andreas Wilke, and Olga Zagnitko. 2008. 'The RAST Server: Rapid Annotations using Subsystems Technology', *BMC Genomics*, 9: 75-75.

Bacosa, Hernando P., Deana L. Erdner, Brad E. Rosenheim, Prateek Shetty, Kiley W. Seitz, Brett J. Baker, and Zhanfei Liu. 2018. 'Hydrocarbon degradation and response of seafloor sediment bacterial community in the northern Gulf of Mexico to light Louisiana sweet crude oil', *The ISME Journal*, 12: 2532-43.

Bagby, Sarah C., Christopher M. Reddy, Christoph Aeppli, G. Burch Fisher, and David L. Valentine. 2017. 'Persistence and biodegradation of oil at the ocean floor following Deepwater Horizon', *Proceedings of the National Academy of Sciences*, 114: E9.

Bairoch, A., and R. Apweiler. 2000. 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Research*, 28: 45-48.

Barron, Mace G. 2011. 'Ecological Impacts of the Deepwater Horizon Oil Spill: Implications for Immunotoxicity', *Toxicologic Pathology*, 40: 315-20.

Beazley, Melanie J., Robert J. Martinez, Suja Rajan, Jessica Powell, Yvette M. Piceno, Lauren M. Tom, Gary L. Andersen, Terry C. Hazen, Joy D. Van Nostrand, Jizhong Zhou, Behzad Mortazavi, and Patricia A. Sobecky. 2012. 'Microbial Community Analysis of a Coastal Salt Marsh Affected by the Deepwater Horizon Oil Spill', *PLOS ONE*, 7: e41305.

Bociu, Ioana, Boryoung Shin, Wm Brian Wells, Joel E. Kostka, Konstantinos T. Konstantinidis, and Markus Huettel. 2019. 'Decomposition of sediment-oil-agglomerates in a Gulf of Mexico sandy beach', *Scientific Reports*, 9: 10071.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30: 2114-20.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, Jessica Jarett, Adam R. Rivers, Emiley A. Eloe-Fadrosh, Susannah G. Tringe, Natalia N. Ivanova, Alex Copeland, Alicia Clum, Eric D. Becraft, Rex R. Malmstrom, Bruce Birren, Mircea Podar, Peer Bork, George M. Weinstock, George M. Garrity, Jeremy A. Dodsworth, Shibu Yooseph, Granger Sutton, Frank O. Glöckner, Jack A. Gilbert, William C. Nelson, Steven J. Hallam, Sean P. Jungbluth, Thijs J. G. Ettema, Scott Tighe, Konstantinos T. Konstantinidis, Wen-Tso Liu, Brett J. Baker, Thomas Rattei, Jonathan A. Eisen, Brian Hedlund, Katherine D. McMahon, Noah Fierer, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Gene W. Tyson, Christian Rinke, Consortium The Genome Standards, Nikos C. Kyrpides, Lynn Schriml, George M. Garrity, Philip Hugenholtz, Granger Sutton, Pelin Yilmaz, Folker Meyer, Frank O. Glöckner, Jack A. Gilbert, Rob Knight, Rob Finn, Guy Cochrane, Ilene Karsch-Mizrachi, Alla Lapidus, Folker Meyer, Pelin Yilmaz, Donovan H. Parks, A. Murat Eren, Lynn Schriml, Jillian F. Banfield, Philip Hugenholtz, and Tanja Woyke. 2017. 'Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea', *Nature Biotechnology*, 35: 725.

Broman, Elias, Johanna Sjöstedt, Jarone Pinhassi, and Mark Dopson. 2017. 'Shifts in coastal sediment oxygenation cause pronounced changes in microbial community composition and associated metabolism', *Microbiome*, 5: 96.

Burdige, David J. 2006. *Geochemistry of marine sediments* (Princeton University Press).

Burdige, David J. 2007. 'Preservation of Organic Matter in Marine Sediments: Controls, Mechanisms, and an Imbalance in Sediment Organic Carbon Budgets?', *Chemical Reviews*, 107: 467-85.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. 'BLAST+: architecture and applications', *BMC Bioinformatics*, 10: 421.

Camilli, Richard, Christopher M. Reddy, Dana R. Yoerger, Benjamin A. S. Van Mooy, Michael V. Jakuba, James C. Kinsey, Cameron P. McIntyre, Sean P. Sylva, and James V. Maloney. 2010. 'Tracking Hydrocarbon Plume Transport and Biodegradation at Deepwater Horizon', *Science*, 330: 201.

Capone, DG. 1993. 'Determination of nitrogenase activity in aquatic samples using the acetylene reduction procedure', *Handbook of methods in aquatic microbial ecology*: 621-31.

Castro, J. C., R. Lm Rodriguez, W. T. Harvey, M. R. Weigand, J. K. Hatt, M. Q. Carter, and K. T. Konstantinidis. 2018. 'imGLAD: accurate detection and quantification of target organisms in metagenomes', *PeerJ*, 6: e5882.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. 'Ribosomal Database Project: data and tools for high throughput rRNA analysis', *Nucleic Acids Research*, 42: D633-D42.

Cox, Murray P., Daniel A. Peterson, and Patrick J. Biggs. 2010. 'SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data', *BMC Bioinformatics*, 11: 485.

Edgar, R. C. 2010. 'Search and clustering orders of magnitude faster than BLAST', *Bioinformatics*, 26: 2460-1.

Ehrenhauss, Sandra, and Markus Huettel. 2004. 'Advective transport and decomposition of chain-forming planktonic diatoms in permeable sediments', *Journal of Sea Research*, 52: 179-97.

Elango, V., M. Urbano, K. R. Lemelle, and J. H. Pardue. 2014. 'Biodegradation of MC252 oil in oil:sand aggregates in a coastal headland beach environment', *Front Microbiol*, 5: 161.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. 'Anvi'o: an advanced analysis and visualization platform for 'omics data', *PeerJ*, 3: e1319.

Forster, Stefan, Markus Huettel, and Wiebke Ziebis. 1996. 'Impact of boundary layer flow velocity on oxygen utilisation in coastal sediments', *Marine Ecology Progress Series*, 143: 173-85.

Galili, Tal. 2015. 'dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering', *Bioinformatics*, 31: 3718-20.

Gibson, David T., and Rebecca E. Parales. 2000. 'Aromatic hydrocarbon dioxygenases in environmental biotechnology', *Current Opinion in Biotechnology*, 11: 236-43.

Guibert, Lilian M., Claudia L. Loviso, Magalí S. Marcos, Marta G. Commendatore, Hebe M. Dionisi, and Mariana Lozada. 2012. 'Alkane Biodegradation Genes from Chronically Polluted Subantarctic Coastal Sediments and Their Shifts in Response to Oil Exposure', *Microbial Ecology*, 64: 605-16.

Hazen, Terry C., Eric A. Dubinsky, Todd Z. DeSantis, Gary L. Andersen, Yvette M. Piceno, Navjeet Singh, Janet K. Jansson, Alexander Probst, Sharon E. Borglin, Julian L. Fortney, William T. Stringfellow, Markus Bill, Mark E. Conrad, Lauren M. Tom, Krystle L. Chavarria, Thana R. Alusi, Regina Lamendella, Dominique C. Joyner, Chelsea Spier, Jacob Baelum, Manfred Auer, Marcin L. Zemla, Romy Chakraborty, Eric L. Sonnenthal, Patrik D'haeseleer, Hoi-Ying N. Holman, Shariff Osman, Zhenmei Lu, Joy D. Van Nostrand, Ye Deng, Jizhong Zhou, and Olivia U. Mason. 2010. 'Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria', *Science*, 330: 204.

Head, Ian M., D. Martin Jones, and Wilfred F. M. Röling. 2006. 'Marine microorganisms make a meal of oil', *Nature Reviews Microbiology*, 4: 173.

Hu, Ping, Eric A. Dubinsky, Alexander J. Probst, Jian Wang, Christian M. K. Sieber, Lauren M. Tom, Piero R. Gardinali, Jillian F. Banfield, Ronald M. Atlas, and Gary L. Andersen. 2017. 'Simulation of Deepwater Horizon; oil plume reveals substrate specialization within a complex community of hydrocarbon degraders', *Proceedings of the National Academy of Sciences*, 114: 7432.

Huettel, Markus, Will A. Overholt, Joel E. Kostka, Christopher Hagan, John Kaba, Wm Brian Wells, and Stacia Dudley. 2018. 'Degradation of Deepwater Horizon oil buried in a Florida beach influenced by tidal pumping', *Marine Pollution Bulletin*, 126: 488-500.

Huettel, Markus, and Antje Rusch. 2000. 'Transport and degradation of phytoplankton in permeable sediment', *Limnology and Oceanography*, 45: 534-49.

Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas, and Jillian F. Banfield. 2016. 'A new view of the tree of life', *Nature Microbiology*, 1: 16048.

Hull, Natalie M., Fangqiong Ling, Ameet J. Pinto, Mads Albertsen, H. Grace Jang, Pei-Ying Hong, Konstantinos T. Konstantinidis, Mark LeChevallier, Rita R. Colwell, and Wen-Tso Liu. 2019. 'Drinking Water Microbiome Project: Is it Time?', *Trends in Microbiology*, 27: 670-77.

Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics*, 11: 119-19.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. 'High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries', *Nature Communications*, 9: 5114.

Ji, Yurui, Guannan Mao, Yingying Wang, and Mark Bartlam. 2013. 'Structural insights into diversity and n-alkane biodegradation mechanisms of alkane hydroxylases', *Frontiers in Microbiology*, 4: 58.

Johnston, Eric R., Luis M. Rodriguez-R, Chengwei Luo, Mengting M. Yuan, Liyou Wu, Zhili He, Edward A. G. Schuur, Yiqi Luo, James M. Tiedje, Jizhong Zhou, and Konstantinos T. Konstantinidis. 2016. 'Metagenomics Reveals Pervasive Bacterial

Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem', *Frontiers in Microbiology*, 7: 579.

Jorgensen, Bo Barker. 1994. "Diffusion processes and boundary layers in microbial mats." In *Microbial Mats*, edited by Lucas J. Stal and Pierre Caumette, 243-53. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jørgensen, Bo Barker. 1982. 'Mineralization of organic matter in the sea bed—the role of sulphate reduction', *Nature*, 296: 643-45.

Karsenti, E., S. G. Acinas, P. Bork, C. Bowler, C. De Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J. M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, and P. Wincker. 2011. 'A holistic approach to marine eco-systems biology', *PLoS Biol*, 9: e1001177.

Karthikeyan, Smruthi, Luis M. Rodriguez-R, Patrick Heritier-Robbins, Janet K. Hatt, Markus Huettel, Joel E. Kostka, and Konstantinos T. Konstantinidis. 2019. 'Genome Repository of Oiled Systems (GROS): an interactive and searchable database that expands the catalogued diversity of crude oil-associated microbes', *bioRxiv*: 838573.

Karthikeyan, Smruthi, Luis M. Rodriguez-R, Patrick Heritier-Robbins, Minjae Kim, Will A. Overholt, John C. Gaby, Janet K. Hatt, Jim C. Spain, Ramon Rosselló-Móra, Markus Huettel, Joel E. Kostka, and Konstantinos T. Konstantinidis. 2019. '"Candidatus Macondimonas diazotrophica", a novel gammaproteobacterial genus dominating crude-oil-contaminated coastal sediments', *The ISME Journal*, 13: 2129-34.

Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*, 30: 3059-66.

Keegan, K. P., E. M. Glass, and F. Meyer. 2016. 'MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function', *Methods Mol Biol*, 1399: 207-33.

Kim, Minjae, Michael R. Weigand, Seungdae Oh, Janet K. Hatt, Raj Krishnan, Ulas Tezel, Spyros G. Pavlostathis, and Konstantinos T. Konstantinidis. 2018. 'Widely

Used Benzalkonium Chloride Disinfectants Can Promote Antibiotic Resistance', *Applied and Environmental Microbiology*, 84: e01201-18.

Kimes, N. E., A. V. Callaghan, D. F. Aktas, W. L. Smith, J. Sunner, B. Golding, M. Drozdowska, T. C. Hazen, J. M. Suflita, and P. J. Morris. 2013. 'Metagenomic analysis and metabolite profiling of deep-sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill', *Front Microbiol*, 4: 50.

Kimes, Nikole E., Amy V. Callaghan, Joseph M. Suflita, and Pamela J. Morris. 2014. 'Microbial transformation of the Deepwater Horizon oil spill—past, present, and future perspectives', *Frontiers in Microbiology*, 5: 603.

King, G. M., J. E. Kostka, T. C. Hazen, and P. A. Sobecky. 2015. 'Microbial responses to the Deepwater Horizon oil spill: from coastal wetlands to the deep sea', *Ann Rev Mar Sci*, 7: 377-401.

Kleindienst, Sara, Sharon Grim, Mitchell Sogin, Annalisa Bracco, Melitza Crespo-Medina, and Samantha B. Joye. 2015. 'Diverse, rare microbial taxa responded to the Deepwater Horizon deep-sea hydrocarbon plume', *The ISME Journal*, 10: 400.

Kleindienst, Sara, Florian-Alexander Herbst, Marion Stagars, Frederick von Netzer, Martin von Bergen, Jana Seifert, Jörg Peplies, Rudolf Amann, Florin Musat, Tillmann Lueders, and Katrin Knittel. 2014. 'Diverse sulfate-reducing bacteria of the Desulfosarcina/Desulfococcus clade are the key alkane degraders at marine seeps', *The ISME Journal*, 8: 2029-44.

Kleindienst, Sara, Michael Seidel, Kai Ziervogel, Sharon Grim, Kathy Loftis, Sarah Harrison, Sairah Y. Malkin, Matthew J. Perkins, Jennifer Field, Mitchell L. Sogin, Thorsten Dittmar, Uta Passow, Patricia M. Medeiros, and Samantha B. Joye. 2015. 'Chemical dispersants can suppress the activity of natural oil-degrading microorganisms', *Proceedings of the National Academy of Sciences*, 112: 14900.

Konstantinidis, Konstantinos T., and Edward F. DeLong. 2008. 'Genomic patterns of recombination, clonal divergence and environment in marine microbial populations', *The ISME Journal*, 2: 1052-65.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. 'Genomic insights that advance the species definition for prokaryotes', *Proceedings of the National Academy of Sciences of the United States of America*, 102: 2567.

Kopylova, Evguenia, Laurent Noé, and Hélène Touzet. 2012. 'SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data', *Bioinformatics*, 28: 3211-17.

Kostka, Joel E., Will A. Overholt, Luis M. Rodriguez-R, Markus Huettel, and Kostas Konstantinidis. 2020. 'Toward a Predictive Understanding of the Benthic Microbial Community Response to Oiling on the Northern Gulf of Mexico Coast.' in Steven A. Murawski, Cameron H. Ainsworth, Sherryl Gilbert, David J. Hollander, Claire B. Paris, Michael Schlüter and Dana L. Wetzel (eds.), *Scenarios and Responses to Future Deep Oil Spills: Fighting the Next War* (Springer International Publishing: Cham).

Kostka, Joel E., Om Prakash, Will A. Overholt, Stefan J. Green, Gina Freyer, Andy Canion, Jonathan Delgardio, Nikita Norton, Terry C. Hazen, and Markus Huettel. 2011. 'Hydrocarbon-Degrading Bacteria and the Bacterial Community Response in Gulf of Mexico Beach Sands Impacted by the Deepwater Horizon Oil Spill', *Appl Environ Microbiol*, 77: 7962-74.

Kuczynski, Justin, Jesse Stombaugh, William Anton Walters, Antonio González, J. Gregory Caporaso, and Rob Knight. 2011. 'Using QIIME to analyze 16S rRNA gene sequences from microbial communities', *Current protocols in bioinformatics*, Chapter 10: Unit10.7-10.7.

Lagesen, Karin, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W. Ussery. 2007. 'RNAmmer: consistent and rapid annotation of ribosomal RNA genes', *Nucleic Acids Res*, 35: 3100-08.

Lamendella, Regina, Steven Strutt, Sharon Borglin, Romy Chakraborty, Neslihan Tas, Olivia Mason, Jenni Hultman, Emmanuel Prestat, Terry Hazen, and Janet Jansson. 2014. 'Assessment of the Deepwater Horizon oil spill impact on Gulf coast microbial communities', *Frontiers in Microbiology*, 5: 130.

Langmead, Ben, and Steven L. Salzberg. 2012. 'Fast gapped-read alignment with Bowtie 2', *Nature methods*, 9: 357-59.

Lea-Smith, David J., Steven J. Biller, Matthew P. Davey, Charles A. R. Cotton, Blanca M. Perez Sepulveda, Alexandra V. Turchyn, David J. Scanlan, Alison G. Smith, Sallie W. Chisholm, and Christopher J. Howe. 2015. 'Contribution of cyanobacterial alkane production to the ocean hydrocarbon cycle', *Proceedings of the National Academy of Sciences*, 112: 13591.

Leahy, J. G., and R. R. Colwell. 1990. 'Microbial degradation of hydrocarbons in the environment', *Microbiological reviews*, 54: 305-15.

Letunic, Ivica, and Peer Bork. 2016. 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic Acids Research*, 44: W242-W45.

Li, Liu, Xueqian Liu, Wen Yang, Feng Xu, Wei Wang, Lu Feng, Mark Bartlam, Lei Wang, and Zihe Rao. 2008. 'Crystal Structure of Long-Chain Alkane Monooxygenase (LadA) in Complex with Coenzyme FMN: Unveiling the Long-Chain Alkane Hydroxylase', *Journal of Molecular Biology*, 376: 453-65.

Li, W., and A. Godzik. 2006. 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences', *Bioinformatics*, 22: 1658-9.

Liu, Zhanfei, Jiqing Liu, Qingzhi Zhu, and Wei Wu. 2012. 'The weathering of oil after theDeepwater Horizonoil spill: insights from the chemical composition of the oil from the sea surface, salt marshes and sediments', *Environmental Research Letters*, 7: 035302.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15: 550.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. 2004. 'ARB: a software environment for sequence data', *Nucleic Acids Res*, 32: 1363-71.

Luo, Chengwei, Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. 2014. 'MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences', *Nucleic Acids Research*, 42: e73-e73.

Luo, Jian, Zohre Kurt, Deyi Hou, and Jim C. Spain. 2015. 'Modeling Aerobic Biodegradation in the Capillary Fringe', *Environ Sci Technol*, 49: 1501-10.

Maddison, Wayne P., and L. Lacey Knowles. 2006. 'Inferring Phylogeny Despite Incomplete Lineage Sorting', *Systematic Biology*, 55: 21-30.

Mason, Olivia U., Terry C. Hazen, Sharon Borglin, Patrick S. G. Chain, Eric A. Dubinsky, Julian L. Fortney, James Han, Hoi-Ying N. Holman, Jenni Hultman, Regina Lamendella, Rachel Mackelprang, Stephanie Malfatti, Lauren M. Tom, Susannah G. Tringe, Tanja Woyke, Jizhong Zhou, Edward M. Rubin, and Janet K. Jansson. 2012. 'Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill', *The ISME Journal*, 6: 1715.

Mason, Olivia U., Nicole M. Scott, Antonio Gonzalez, Adam Robbins-Pianka, Jacob Bælum, Jeffrey Kimbrel, Nicholas J. Bouskill, Emmanuel Prestat, Sharon Borglin, Dominique C. Joyner, Julian L. Fortney, Diogo Jurelevicius, William T. Stringfellow, Lisa Alvarez-Cohen, Terry C. Hazen, Rob Knight, Jack A. Gilbert, and Janet K. Jansson. 2014. 'Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill', *The ISME Journal*, 8: 1464.

McGenity, Terry J. 2014. 'Hydrocarbon biodegradation in intertidal wetland sediments', *Current Opinion in Biotechnology*, 27: 46-54.

Meckenstock, R. U., E. Annweiler, W. Michaelis, H. H. Richnow, and B. Schink. 2000. 'Anaerobic naphthalene degradation by a sulfate-reducing enrichment culture', *Applied and Environmental Microbiology*, 66: 2743-47.

Momper, Lily, Sean P. Jungbluth, Michael D. Lee, and Jan P. Amend. 2017. 'Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community', *The ISME Journal*, 11: 2319-33.

Moreno-Hagelsieb, Gabriel, and Kristen Latimer. 2007. 'Choosing BLAST options for better detection of orthologs as reciprocal best hits', *Bioinformatics*, 24: 319-24.

Mortazavi, Behzad, Agota Horel, Melanie J. Beazley, and Patricia A. Sobecky. 2013. 'Intrinsic rates of petroleum hydrocarbon biodegradation in Gulf of Mexico intertidal sandy sediments and its enhancement by organic substrates', *Journal of Hazardous Materials*, 244-245: 537-44.

Munoz, Raul, Pablo Yarza, and Ramon Rosselló-Móra. 2014. 'Harmonized Phylogenetic Trees for The Prokaryotes.' in Eugene Rosenberg, Edward F. DeLong, Stephen Lory, Erko Stackebrandt and Fabiano Thompson (eds.), *The Prokaryotes: Actinobacteria* (Springer Berlin Heidelberg: Berlin, Heidelberg).

Murphy, Sarah A., Shengnan Meng, Benson M. Solomon, Dewamunnage M. C. Dias, Timothy J. Shaw, and John L. Ferry. 2016. 'Hydrous Ferric Oxides in Sediment

Catalyze Formation of Reactive Oxygen Species during Sulfide Oxidation', *Frontiers in Marine Science*, 3: 227.

Murray, Karen J., Paul D. Boehm, and Roger C. Prince. 2020. 'The Importance of Understanding Transport and Degradation of Oil and Gasses from Deep-Sea Blowouts.' in Steven A. Murawski, Cameron H. Ainsworth, Sherryl Gilbert, David J. Hollander, Claire B. Paris, Michael Schlüter and Dana L. Wetzel (eds.), *Deep Oil Spills: Facts, Fate, and Effects* (Springer International Publishing: Cham).

Nie, Yong, Chang-Qiao Chi, Hui Fang, Jie-Liang Liang, She-Lian Lu, Guo-Li Lai, Yue-Qin Tang, and Xiao-Lei Wu. 2014. 'Diverse alkane hydroxylase genes in microorganisms and environments', *Scientific Reports*, 4: 4968.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome Biology*, 17: 132.

Orcutt, Beth N., Samantha B. Joye, Sara Kleindienst, Katrin Knittel, Alban Ramette, Anja Reitz, Vladimir Samarkin, Tina Treude, and Antje Boetius. 2010. 'Impact of natural oil and higher hydrocarbons on microbial diversity, distribution, and activity in Gulf of Mexico cold-seep sediments', *Deep Sea Research Part II: Topical Studies in Oceanography*, 57: 2008-21.

Orellana, Luis H., Luis M. Rodriguez-R, and Konstantinos T. Konstantinidis. 2016. 'ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores', *Nucleic Acids Research*, 45: e14-e14.

Ortega-Calvo, J. J., and P. M. Gschwend. 2010. 'Influence of low oxygen tensions and sorption to sediment black carbon on biodegradation of pyrene', *Appl Environ Microbiol*, 76: 4430-7.

Overbeek, Ross, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, Veronika Vonstein, Alice R. Wattam, Fangfang Xia, and Rick Stevens. 2014. 'The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)', *Nucleic Acids Research*, 42: D206-D14.

Overton, Edward B., Terry L. Wade, Jagoš R. Radović, Buffy M. Meyer, M. Scott Miles, and Stephen R. Larter. 2016. 'Chemical Composition of Macondo and Other

Crude Oils and Compositional Alterations During Oil Spills', *Oceanography*, 29: 50-63.

Owens, Edward H., and Gary A. Sergy. 2003. 'The development of the SCAT process for the assessment of oiled shorelines', *Marine Pollution Bulletin*, 47: 415-22.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome research*, 25: 1043-55.

Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. 'Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life', *Nature Microbiology*, 2: 1533-42.

Paul, John H., David Hollander, Paula Coble, Kendra L. Daly, Sue Murasko, David English, Jonelle Basso, Jennifer Delaney, Lauren McDaniel, and Charles W. Kovach. 2013. 'Toxicity and Mutagenicity of Gulf of Mexico Waters During and After the Deepwater Horizon Oil Spill', *Environ Sci Technol*, 47: 9651-59.

Peng, Y., H. C. Leung, S. M. Yiu, and F. Y. Chin. 2012. 'IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth', *Bioinformatics*, 28: 1420-8.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2009. 'FastTree: computing large minimum evolution trees with profiles instead of a distance matrix', *Molecular biology and evolution*, 26: 1641-50.

Prince, R. C. 2010. 'Bioremediation of Marine Oil Spills.' in Kenneth N. Timmis (ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (Springer Berlin Heidelberg: Berlin, Heidelberg).

Prince, R. C., A. Gramain, and T. J. McGenity. 2010. 'Prokaryotic Hydrocarbon Degraders.' in Kenneth N. Timmis (ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (Springer Berlin Heidelberg: Berlin, Heidelberg).

Prince, Roger C., and Ronald M. Atlas. 2018. 'Bioremediation of Marine Oil Spills.' in Robert Steffan (ed.), *Consequences of Microbial Interactions with Hydrocarbons,*

Oils, and Lipids: Biodegradation and Bioremediation (Springer International Publishing: Cham).

Pritchard, P. H., J. G. Mueller, J. C. Rogers, F. V. Kremer, and J. A. Glaser. 1992. 'Oil spill bioremediation: experiences, lessons and results from the Exxon Valdez oil spill in Alaska', *Biodegradation*, 3: 315-35.

Pruesse, E., J. Peplies, and F. O. Glockner. 2012. 'SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes', *Bioinformatics*, 28: 1823-9.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. 2013. 'The SILVA ribosomal RNA gene database project: improved data processing and web-based tools', *Nucleic Acids Res*, 41: D590-6.

Rho, M., H. Tang, and Y. Ye. 2010. 'FragGeneScan: predicting genes in short and error-prone reads', *Nucleic Acids Res*, 38: e191.

Rodriguez, R. Lm, S. Gunturu, W. T. Harvey, R. Rossello-Mora, J. M. Tiedje, J. R. Cole, and K. T. Konstantinidis. 2018. 'The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level', *Nucleic Acids Res*, 46: W282-w88.

Rodriguez, R. Lm, and K. T. Konstantinidis. 2014. 'Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets', *Bioinformatics*, 30: 629-35.

Rodriguez-R, L. M., D. Tsementzi, C. Luo, and K. T. Konstantinidis. 2019. 'Iterative Subtractive Binning of Freshwater Chronoseries Metagenomes Identifies of over Four Hundred Novel Species and their Ecologic Preferences', *bioRxiv*: 826941.

Rodriguez-R, Luis M, and Konstantinos T Konstantinidis. 2016a. "The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes." In.: PeerJ Preprints.

Rodriguez-R, Luis M., Santosh Gunturu, William T. Harvey, Ramon Rosselló-Mora, James M. Tiedje, James R. Cole, and Konstantinos T. Konstantinidis. 2018. 'The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level', *Nucleic Acids Res*, 46: W282-W88.

Rodriguez-R, Luis M., and Konstantinos T. Konstantinidis. 2016b. 'The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes', *PeerJ Preprints*, 4: e1900v1.

Rodriguez-R, Luis M., Will A. Overholt, Christopher Hagan, Markus Huettel, Joel E. Kostka, and Konstantinos T. Konstantinidis. 2015. 'Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill', *The ISME Journal*, 9: 1928-40.

Rojo, Fernando. 2009. 'Degradation of alkanes by bacteria', *Environmental Microbiology*, 11: 2477-90.

Ruddy, Brian M., Markus Huettel, Joel E. Kostka, Vladislav V. Lobodin, Benjamin J. Bythell, Amy M. McKenna, Christoph Aeppli, Christopher M. Reddy, Robert K. Nelson, Alan G. Marshall, and Ryan P. Rodgers. 2014. 'Targeted Petroleomics: Analytical Investigation of Macondo Well Oil Oxidation Products from Pensacola Beach', *Energy & Fuels*, 28: 4043-50.

Schirmer, A., M. A. Rude, X. Li, E. Popova, and S. B. del Cardayre. 2010. 'Microbial biosynthesis of alkanes', *Science*, 329: 559-62.

Schneiker, Susanne, Vítor A. P. Martins dos Santos, Daniela Bartels, Thomas Bekel, Martina Brecht, Jens Buhrmester, Tatyana N. Chernikova, Renata Denaro, Manuel Ferrer, Christoph Gertler, Alexander Goesmann, Olga V. Golyshina, Filip Kaminski, Amit N. Khachane, Siegmund Lang, Burkhard Linke, Alice C. McHardy, Folker Meyer, Taras Nechitaylo, Alfred Pühler, Daniela Regenhardt, Oliver Rupp, Julia S. Sabirova, Werner Selbitschka, Michail M. Yakimov, Kenneth N. Timmis, Frank-Jörg Vorhölter, Stefan Weidner, Olaf Kaiser, and Peter N. Golyshin. 2006. 'Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis', *Nature Biotechnology*, 24: 997-1004.

Seitz, Kiley W., Nina Dombrowski, Laura Eme, Anja Spang, Jonathan Lombard, Jessica R. Sieber, Andreas P. Teske, Thijs J. G. Ettema, and Brett J. Baker. 2019. 'Asgard archaea capable of anaerobic hydrocarbon cycling', *Nature Communications*, 10: 1822.

Sekar, Ramanan, and Thomas J. DiChristina. 2017. 'Degradation of the recalcitrant oil spill components anthracene and pyrene by a microbially driven Fenton reaction', *FEMS Microbiology Letters*, 364.

Shanklin, John, and Edward Whittle. 2003. 'Evidence linking the Pseudomonas oleovorans alkane ω-hydroxylase, an integral membrane diiron enzyme, and the fatty acid desaturase family', *FEBS Letters*, 545: 188-92.

Shao, Zongze, and Wanpeng Wang. 2013. 'Enzymes and genes involved in aerobic alkane degradation', *Frontiers in Microbiology*, 4: 116.

Shin, Boryoung, Minjae Kim, Karsten Zengler, Kuk-Jeong Chin, Will A. Overholt, Lisa M. Gieg, Konstantinos T. Konstantinidis, and Joel E. Kostka. 2019. 'Anaerobic degradation of hexadecane and phenanthrene coupled to sulfate reduction by enriched consortia from northern Gulf of Mexico seafloor sediment', *Scientific Reports*, 9: 1239.

Short, Jeffrey W., Mandy R. Lindeberg, Patricia M. Harris, Jacek M. Maselko, Jerome J. Pella, and Stanley D. Rice. 2004. 'Estimate of Oil Persisting on the Beaches of Prince William Sound 12 Years after the Exxon Valdez Oil Spill', *Environ Sci Technol*, 38: 19-25.

Sievers, Fabian, and Desmond G. Higgins. 2018. 'Clustal Omega for making accurate alignments of many protein sequences', *Protein science : a publication of the Protein Society*, 27: 135-45.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. 2011. 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology*, 7.

Smith, C. B., B. B. Tolar, J. T. Hollibaugh, and G. M. King. 2013. 'Alkane hydroxylase gene (alkB) phylotype composition and diversity in northern Gulf of Mexico bacterioplankton', *Front Microbiol*, 4: 370.

Stamatakis, Alexandros. 2014. 'RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*, 30: 1312-13.

Su, Xiaoquan, Weihua Pan, Baoxing Song, Jian Xu, and Kang Ning. 2014. 'Parallel-META 2.0: Enhanced Metagenomic Data Analysis with Functional Annotation, High Performance Computing and Advanced Visualization', *PLOS ONE*, 9: e89323.

Sun, Shaojie, Chuanmin Hu, Oscar Garcia-Pineda, Vassiliki Kourafalou, Matthieu Le Hénaff, and Yannis Androulidakis. 2018. 'Remote sensing assessment of oil spills near a damaged platform in the Gulf of Mexico', *Marine Pollution Bulletin*, 136: 141-51.

Teske, Andreas. 2019. 'Hydrocarbon-degrading microbial communities in natural oil seeps', *Microbial Communities Utilizing Hydrocarbons and Lipids: Members, Metagenomics and Ecophysiology*: 1-31.

Throne-Holst, Mimmi, Alexander Wentzel, Trond E. Ellingsen, Hans-Kristian Kotlar, and Sergey B. Zotchev. 2007. 'Identification of Novel Genes Involved in Long-Chain Alkane Degradation by Acinetobacter sp. Strain DSM 17874', *Applied and Environmental Microbiology*, 73: 3327.

Tsementzi, Despina, Luis M. Rodriguez-R, Carlos A. Ruiz-Perez, Alexandra Meziti, Janet K. Hatt, and Konstantinos T. Konstantinidis. 2019. '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*, 42: 495-505.

Urakawa, Hidetoshi, Suja Rajan, Megan E. Feeney, Patricia A. Sobecky, and Behzad Mortazavi. 2019. 'Ecological response of nitrification to oil spills and its impact on the nitrogen cycle', *Environmental Microbiology*, 21: 18-33.

Valentine, David L., John D. Kessler, Molly C. Redmond, Stephanie D. Mendes, Monica B. Heintz, Christopher Farwell, Lei Hu, Franklin S. Kinnaman, Shari Yvon-Lewis, Mengran Du, Eric W. Chan, Fenix Garcia Tigreros, and Christie J. Villanueva. 2010. 'Propane Respiration Jump-Starts Microbial Response to a Deep Oil Spill', *Science*, 330: 208.

Valentine, David L., and Christopher M. Reddy. 2015. 'Latent hydrocarbons from cyanobacteria', *Proceedings of the National Academy of Sciences*, 112: 13434.

Venosa, Albert D., and Xueqing Zhu. 2003. 'Biodegradation of Crude Oil Contaminating Marine Shorelines and Freshwater Wetlands', *Spill Science & Technology Bulletin*, 8: 163-78.

Wakeham, Stuart G., and Elizabeth A. Canuel. 2006. 'Degradation and Preservation of Organic Matter in Marine Sediments.' in John K. Volkman (ed.), *Marine Organic Matter: Biomarkers, Isotopes and DNA* (Springer Berlin Heidelberg: Berlin, Heidelberg).

Wang, Liping, Wanpeng Wang, Qiliang Lai, and Zongze Shao. 2010. 'Gene diversity of CYP153A and AlkB alkane hydroxylases in oil-degrading bacteria isolated from the Atlantic Ocean', *Environmental Microbiology*, 12: 1230-42.

Wasmund, Kenneth, Marc Mußmann, and Alexander Loy. 2017. 'The life sulfuric: microbial ecology of sulfur cycling in marine sediments', *Environmental Microbiology Reports*, 9: 323-44.

White, Helen K., Pen-Yuan Hsing, Walter Cho, Timothy M. Shank, Erik E. Cordes, Andrea M. Quattrini, Robert K. Nelson, Richard Camilli, Amanda W. J. Demopoulos, Christopher R. German, James M. Brooks, Harry H. Roberts, William Shedd, Christopher M. Reddy, and Charles R. Fisher. 2012. 'Impact of the Deepwater Horizon oil spill on a deep-water coral community in the Gulf of Mexico', *Proceedings of the National Academy of Sciences*, 109: 20303.

White, Helen K., Charles T. Marx, David L. Valentine, Charles Sharpless, Christoph Aeppli, Kelsey M. Gosselin, Veronika Kivenson, Rachel M. Liu, Robert K. Nelson, Sean P. Sylva, and Christopher M. Reddy. 2019. 'Examining Inputs of Biogenic and Oil-Derived Hydrocarbons in Surface Waters Following the Deepwater Horizon Oil Spill', *ACS Earth and Space Chemistry*, 3: 1329-37.

Widdel, F. 2010. 'Cultivation of Anaerobic Microorganisms with Hydrocarbons as Growth Substrates.' in Kenneth N. Timmis (ed.), *Handbook of Hydrocarbon and Lipid Microbiology* (Springer Berlin Heidelberg: Berlin, Heidelberg).

Widdel, F., and R. Rabus. 2001a. 'Anaerobic biodegradation of saturated and aromatic hydrocarbons', *Curr Opin Biotechnol*, 12: 259-76.

Widdel, Friedrich, and Friedhelm Bak. 1992. 'Gram-Negative Mesophilic Sulfate-Reducing Bacteria.' in Albert Balows, Hans G. Trüper, Martin Dworkin, Wim Harder and Karl-Heinz Schleifer (eds.), *The Prokaryotes: A Handbook on the Biology of Bacteria: Ecophysiology, Isolation, Identification, Applications* (Springer New York: New York, NY).

Widdel, Friedrich, and Ralf Rabus. 2001b. 'Anaerobic biodegradation of saturated and aromatic hydrocarbons', *Current Opinion in Biotechnology*, 12: 259-76.

Wu, Y. W., B. A. Simmons, and S. W. Singer. 2016. 'MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets', *Bioinformatics*, 32: 605-7.

Yarza, P., W. Ludwig, J. Euzeby, R. Amann, K. H. Schleifer, F. O. Glockner, and R. Rossello-Mora. 2010. 'Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses', *Syst Appl Microbiol*, 33: 291-9.

Zaremba-Niedzwiedzka, Katarzyna, Eva F. Caceres, Jimmy H. Saw, Disa Bäckström, Lina Juzokaite, Emmelien Vancaester, Kiley W. Seitz, Karthik Anantharaman, Piotr Starnawski, Kasper U. Kjeldsen, Matthew B. Stott, Takuro Nunoura, Jillian F. Banfield, Andreas Schramm, Brett J. Baker, Anja Spang, and Thijs J. G. Ettema. 2017. 'Asgard archaea illuminate the origin of eukaryotic cellular complexity', *Nature*, 541: 353.

Zhang, Jiajie, Kassian Kobert, Tomáš Flouri, and Alexandros Stamatakis. 2014. 'PEAR: a fast and accurate Illumina Paired-End reAd mergeR', *Bioinformatics*, 30: 614-20.

Zhao, Shilin, Yan Guo, Quanhu Sheng, and Yu Shyr. 2014. 'Heatmap3: an improved heatmap package with more powerful and convenient features', *BMC Bioinformatics*, 15: P16.

Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. 'Ab initio gene identification in metagenomic sequences', *Nucleic Acids Res*, 38: e132.