The metagenomic signatures of impacted environments: Unravelling the microbial community dynamics in ecosystem function

Renee J. Smith

BSc Hons

Thesis submitted for the degree of Doctor of Philosophy

September 2012

School of Biological Sciences

Flinders University

Adelaide, Australia

Table of Contents

Summary

Microbes are largely responsible for the turnover of energy and matter and are thus, integral players in ecosystem functioning. Despite the increasing awareness of the importance of microbial communities, there is still a critical lack of information on the complex relationship between microbial communities and the environment. Metagenomic analysis is thought to yield the most quantitative and accurate view of the microbial world, greatly increasing our ability to generate microbial profiles of the changing world. These methodologies have led to the growing interest in understanding and forecasting microbial responses to anthropogenic disturbances. This thesis investigates the microbial responses to two common forms of pollution, agricultural modification and hydrocarbon impact, to determine to what extent the resident microbial communities may be effected by introduced contaminants. The reoccurring theme of this thesis has been that major shifts in the structure and function of the resident microbial communities was observed following environmental change. Moreover, this thesis demonstrated that the microbial communities inhabiting impacted environments exhibited markedly different community responses based on contaminant type, allowing for the discrimination of their metagenomic signatures. This thesis provides detailed insight into how environmental change affects the inhabiting microbial consortia, and for the first time, demonstrates how the overall metagenomic signature can be used to detect and assess the extent to which anthropogenic disturbances have altered our planet.

Acknowledgements

I would firstly like that thank my primary supervisor Assoc. Prof. Jim Mitchell for his mentoring, guidance and support throughout this project. Thank you for believing in me (even when I didn't), the humour that kept me sane and above all else, the opportunities that allowed me to be the best that I could be. I truly appreciate everything you have done for me. I would also like to thank my co-supervisor Assoc. Prof. Melissa Brown for your encouragement and advice over the course of this study.

A big thanks to the 'Mitchell-Seuront Lab' over the years, in particular to Tom Jeffries, your endless patience and assistance throughout my time in this lab has helped me more than you could possibly know. To Kelly Newton, Ben Roudnew, Justin Seymour and James Paterson, thank-you for your advice, training, support and friendship from the start of this project. I would also like to thank members of the Evolutionary Biology Unit, in particular Alison Fitch who has been a great source of information and support during the many, many hours of trial and errors in the lab.

Also thank you to my friends and family, in particular to my parents, who have been a constant source of inspiration for me. I would not be where I am today if it was not for your constant love, support and encouragement, and for that I am forever grateful.

And last but not least I would like to thank my partner Michael. You have been amazing throughout the course of this PhD and I am constantly surprised by your generosity, humour, friendship and love. Without you I could not have come this far.

Declaration

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Renee Jade Smith

September 2012

Chapter 1

General Introduction

1.1 Microbial communities run the world

Microorganisms are the most abundant and diverse group on the planet, with estimates of 4-6 × 10^{30} prokaryotic cells on earth (Whitman *et al*., 1998; van der Heijden *et al*., 2008; DeLong, 2009). Although invisible to the naked eye, microbes are ubiquitous, diverse and essential components of all ecosystems (Whitman *et al*., 1998; DeLong and Pace, 2001; Fraser *et al*., 2009). This is largely due to their fundamental role in the turnover of energy and matter, subsequently forming the basis of environmental food webs (Steele *et al*., 2011). For example, microbial communities are known to convert carbon, nitrogen, oxygen and sulfur into forms accessible to all other living things (Whitman *et al*., 1998; Karl, 2002; Rittman *et al*., 2008). Microbes are also heavily relied upon for the degradation and clean-up of pollutants in the environment (Hemme *et al*., 2010; Kostka *et al*., 2011; Liang *et al*., 2011). These processes are all achieved by complex microbial networks, which have the capacity to adapt to, and transform the world around them (Follows *et al*., 2007; Lawrence *et al*., 2012). Due to these capabilities, ecosystem functioning and microbial communities are intimately connected (Chapin III *et al*., 1997; Gianoulis *et al*., 2009).

Despite their importance to ecosystem functioning, microbes remain largely unknown, with current estimates of the diversity of microbial life being at least 100 times greater than previously thought (Sogin *et al*., 2006; Kunin *et al*., 2010). The breadth and newness of this diversity means that the complex relationships between microbial community composition and the environment are still being decoded (Zengler and Palsson, 2012). This gap in knowledge is largely due to methodological limitations as well as their overwhelming diversity and abundance (Woyke *et al*., 2009; Maron *et al*., 2011; Martinez-Garcia *et al*., 2012). Advances in metagenomic

sequencing technologies, however, have allowed for the direct sequencing of representative segments of whole environmental microbial communities, greatly increasing our ability to generate microbial profiles of environmental systems (Wommack *et al*., 2008; Kennedy *et al*., 2010; Xing *et al*., 2012). Combining these high throughput sequencing methods with computational tools such as multivariate analysis, could then provide insight into the tracking, manipulation and discrimination of microbial communities (Gonzalez *et al*., 2012). Consequently, this has led to the growing interest in forecasting and understanding microbiological responses to anthropogenic disturbances on all scales (Barnosky *et al*., 2012), with a special focus on the microbial communities (Ager *et al*., 2010; Berga *et al*., 2012).

1.2 Microbial communities as biological indicators

Baas-Becking and Beijerink (Bass Becking, 1934; de Wit and Bouvier, 2006) hypothesized that microbial taxa have preferred environments: “Everything is everywhere, but the environment selects.” In other words, microorganisms are ubiquitously dispersed globally, however, unique environmental conditions, as well as the microbes functional capabilities, determine their dominance (Keller and Hettich, 2009). There is dispute about the idea that “everything is everywhere”, with recent evidence of the global occurrence and geographically localised occurrence of some microbial species (Ramette and Tiedje, 2007; Zinger *et al*., 2011). However, pollution events have been shown to leave lasting signatures on microbial assemblages that are evident at distances as small as 500 km, generating evidence to support the theory that different contemporary environments maintain distinctive microbial assemblages (Martiny *et al*., 2006; Jeffries *et al*., 2011a; Marchetti *et al*., 2012).

It is therefore, not surprising that changes to the surrounding environment can lead to a major shift in the structure and function of the microbial consortia (Dinsdale *et al*., 2008a; Hemme *et al*., 2010; Jeffries *et al*., 2011b). Once these shifts in structure and function are characterised, microbial community dynamics can be used to predict environmental conditions (Fuhrman *et al*., 2006; Dinsdale *et al*., 2008b; Fuhrman, 2009; Gianoulis *et al*., 2009; Larsen *et al*., 2012). Therefore, understanding the intimate relationship between microbial communities and the factors that control them is particularly important given the increase in anthropogenic activities (Fuhrman *et al*., 2006; Ager *et al*., 2010; Stegen *et al*., 2012).

1.3 Anthropogenic disturbances

Current global environmental disturbances that effect diversity and composition of microbial communities are profoundly altering biosphere functioning (Chapin III *et al*., 1997; Balser *et al*., 2006; Sjöstedt *et al*., 2012). Among the disturbances threatening ecosystem health globally are agricultural modification and pollution events (Ager *et al*., 2010; Malone *et al*., 2010; Carpenter *et al*., 2011). For example, it has been estimated that approximately 40% of land surface has been converted for agricultural practises, becoming one of the largest terrestrial biomes on the planet (Asner *et al*., 2004; Foley *et al*., 2005; Lee *et al*., 2011). Furthermore, long term effects have been associated with agriculturally influenced land, whereby fields that have been abandoned for nine years still exhibited similar microbial community compositions when compared to actively cultivated land (Buckley and Schmidt, 2003). Therefore, it is now widely accepted that agricultural practises can dramatically change microbial community dynamics and thus, ecosystem functioning (Mäder *et al*., 2002; Kaye *et al*., 2005; Ge *et al*., 2008; Sun *et al*., 2011).

The effects of hydrocarbon impact are also widely studied due to its long term toxicity and persistence worldwide (Vinas *et al*., 2005; Kostka *et al*., 2011; Liang *et al*., 2011). Due to its natural occurrence in the environment, numerous microorganisms have evolved the capability of utilizing hydrocarbons as energy sources (Atlas and Hazen, 2011). Their ability to effectively remediate hydrocarbons in the environment means that microbial communities are commonly used for bioremediation; however, the mechanisms by which this is achieved in the natural environment are still being elucidated (Chakraborty *et al*., 2012). Thus, knowledge about the shifts in microbial community structure and functionality following disturbances could improve our understanding of ecosystem processes, and thus improve management strategies (Mäder *et al*., 2002; Ge *et al*., 2008; Griffiths and Philippot, 2012).

Previous metagenomic studies have shown that contamination can lead to rare taxa or metabolic processes becoming more prominent, thereby linking the community composition to environmental change (Dinsdale *et al*., 2008a; Jeffries *et al*., 2011b). However the majority of these studies have focused on discrete environments effected by a single contaminant. Furthermore, studies have shown that substrate type, for example sediment versus water, drives the overall structure and functionality of an environmental microbial community, over that of the chemical properties (Jeffries *et al*., 2011a). Thus, diverse substrate types, exhibiting different contamination events, provides a means by which metagenomic signatures can be generated to discriminate between impacted microbial communities.

1.4 Thesis Objective

The primary objective of this thesis was to investigate two common forms of pollution, agricultural modification and hydrocarbon impact, from two different environments, groundwater and sediment, respectively. The metagenomic data produced will provide insight into the taxonomy and metabolic processes of the resident microbial communities, and to determine to what extent these may be affected by introduced contaminants.

Specifically the aims were:

1. To determine the impact of agricultural contamination on unconfined aquifer microbial community structure and function, with the goal to find signature community changes indicative of contamination

2. To determine the impact of historical hydrocarbon contamination on the microbial community structure and function in a marine foreshore environment, to provide insight into the signature community changes indicative of contamination.

3. To provide novel insight into the viral community profile in groundwater ecosystems, including the discrimination of any potential pathogens.

4. To determine the extent to which metagenomic signatures can be used to discriminate between contaminant types in impacted environments.

1.5 Thesis Structure

This thesis is formatted in manuscript form for journal submission, each chapter addressing a specific aim. The results from Chapters 2 to 6 are published in peer-reviewed journals, have been submitted for publication, or will be submitted for publication in the near future, thus there is some redundancy in the introduction and methods for each chapter. Chapter 2 assessed the microbial communities residing in unconfined and confined aquifer ecosystems and was published in *Environmental Microbiology* (14: 240-253, 2011). Chapter 3 constructed a viral community profile in the unconfined and confined aquifers in comparison to investigate the survival and spread of viruses in groundwater, and has been submitted for publication in *Environmental Microbiology Reports* (23rd July 2012). Chapter 4 focuses on the indigenous microbial communities inhabiting a historically hydrocarbon impacted beach. Chapter 5 investigates the microbial metabolic footprints associated with hydrocarbon impact, and has been submitted for publication in *PLoS One* (26th July 2012). Chapter 6 elucidates and metagenomic signatures, taxonomic and metabolic, of various introduced contaminants for the potential use as biological indicators. The discussion and implications of these results form Chapter 7. A single reference list has been included at the end of this thesis that includes all literature cited throughout to reduce redundancy.

Chapter 2

Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems

Published as:

Smith RS, Jeffries TC, Roudnew B, Fitch AJ, Seymour JR, Delpin MW, Newton K, Brown MH, Mitchell JG (2011) Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. *Environmental Microbiology* 14: 240-253.

2.0 Summary

A metagenomic analysis of two aquifer systems located under a dairy farming region was performed to examine to what extent the composition and function of microbial communities varies between confined and surface-influenced unconfined groundwater ecosystems. A fundamental shift in taxa was seen with an overrepresentation of *Rhodospirillales*, *Rhodocyclales*, *Chlorobia* and *Circovirus* in the unconfined aquifer, while *Deltaproteobacteria* and *Clostridiales* were overrepresented in the confined aquifer. A relative overrepresentation of metabolic processes including antibiotic resistance (β-lactamase genes), lactose and glucose utilization and DNA replication were observed in the unconfined aquifer, while flagella production, phosphate metabolism and starch uptake pathways were all overrepresented in the confined aquifer. These differences were likely driven by differences in the nutrient status and extent of exposure to contaminants of the two groundwater systems. However, when compared to freshwater, ocean, sediment and animal gut metagenomes, the unconfined and confined aquifers were taxonomically and metabolically more similar to each other than to any other environment. This suggests that intrinsic features of groundwater ecosystems, including low oxygen levels and a lack of sunlight, have provided specific niches for evolution to create unique microbial communities. Obtaining a broader understanding of the structure and function of microbial communities inhabiting different groundwater systems is particularly important given the increased need for managing groundwater reserves of potable water.

2.1 Introduction

Terrestrial subsurface environments, including groundwater, accommodate the largest reservoir of microbes in the biosphere, with estimates of bacterial abundances reaching 3.8-6.0 × 10^{30} cells (Whitman *et al*., 1998). Due to the lack of sunlight and input of nutrients and energy from external sources, these microbial communities are largely responsible for the turnover of energy and matter, forming the basis of subterranean food webs (Sherr and Sherr, 1991). These communities also influence the purity of groundwater and subsequent availability of potable drinking water (Danielopol *et al*., 2003).

Holding more than 97% of the world's freshwater reserves, aquifers are a largely untapped resource of potable drinking water, but also harbour a high diversity of microbes (Gibert and Deharveng, 2002). These reserves are becoming increasingly important (Bond *et al*., 2008) in countries such as Australia, which are susceptible to drought events (Mpelasoka *et al*., 2008). However, the nature of the microbial communities inhabiting aquifers remains largely unexplored. To effectively understand and maintain groundwater reserves it is important to investigate the identity and biogeochemical function of the microbes within aquifer systems.

Aquifer systems, defined by a permeable zone below the earth's surface through which groundwater moves (Hamblin and Christiansen, 2004), are generally classified into two major types; unconfined and confined aquifers. 'Unconfined aquifers' are connected to the surface via open pore space and thus, can receive external input from the surrounding area. They are sensitive to precipitation via seepage through the soil, and are directly affected by human impact (Al-Zabet, 2002). 'Confined aquifers' occur at greater depth and lie below an impermeable strata layer. The thick confining

strata layer ensures that there is no input from the overlaying surface environment. Input to confined aquifers occurs only from distant recharge sources and due to slow flow rates, can be isolated for hundreds to thousands of years (Gibert and Deharveng, 2002). Microbes inhabiting these systems must be capable of surviving with limited resources, as external inputs of nutrients and oxygen are not readily available (Pedersen, 2000; Griebler and Lueders, 2009). Survival strategies to cope in this environment include increased affinity to limiting nutrients and reduced metabolic rates and growth (Teixeira de Mattos and Neijssel, 1997; Brune *et al*., 2000).

Sporadic changes in limiting resources in these groundwater systems, driven by external input, can lead to major shifts in the taxonomy and the metabolism of microbial communities (Hemme *et al*., 2010). The sensitivity of microbes to environmental change allows them to be used as bioindicators (Avidano *et al*., 2005; Steube *et al*., 2009). A major goal in the study of groundwater microbiology is to determine what the effects of these shifts in microbial ecology have on water quality (Langworthy *et al*., 1998; Hemme *et al*., 2010).

The concentration of chemical contaminants and pathogens in groundwater systems is influenced by the biogeochemical and ecological dynamics of subterranean microbial communities (Hemme *et al*., 2010). Shifts in microbial taxonomy resulting from pollution in groundwater have been investigated (Männistö *et al*., 1999; Chang *et al*., 2001) but the effects of introduced contaminants on the metabolic potential of groundwater microbes is only vaguely understood. Previous groundwater studies have shown that microbes respond to external contaminants at both the phenotypic and genotypic level, with changes in microbial community structure, as well as an increase in the number of genes responsible for the degradation of

introduced contaminants (Langworthy *et al*., 1998). Furthermore, Hemme *et al*. (2010) showed that introduced contaminants into groundwater systems can decrease species and allelic diversity and eliminate some metabolic pathways. Evolutionary analysis of a microbial community in groundwater contaminated with heavy metals has shown that lateral gene transfer could play a key role in the rapid response and adaptation to environmental contamination (Hemme *et al*., 2010). Hence, to obtain a complete description of the effect of external influences on groundwater systems, both the taxonomy and the metabolic potential of microbial communities need to be studied.

The effect of agricultural modification on groundwater is less well characterised, however it has been shown that introduced manure from a live-stock farm caused the microbial composition of previously uncontaminated groundwater to taxonomically resemble livestock wastewater (Cho and Kim, 2000). This study used 16S rDNA technology which is limited to prokaryote taxonomy and discounts viruses and eukaryotes. Advances in metagenomic studies have allowed for the direct sequencing of whole environmental microbial genomes (Kennedy *et al*., 2010) and have greatly increased our knowledge of gene function, metabolic processes, community structure and ecosystems response to environmental change. Previous metagenomic studies have revealed clear shifts in the structure of microbial assemblages related to human impact (Dinsdale *et al*., 2008a).

With this in mind, the aim of the present study is to compare an unconfined and a confined groundwater system using metagenomic approaches, and provide insight into the endemic taxonomy and metabolic processes of the resident microbial communities, and how these may be affected by introduced contaminants.

2.2 Results

2.2.1 Overview of the biogeochemical environment and microbial enumeration

The unconfined and confined aquifers were characterised by low oxygen levels of 0.2 mg L^{-1} and 0.26 mg L^{-1} respectively. Iron, sulphur and total organic carbon were all significantly higher ($P < 0.05$) in the unconfined aquifer than the confined aquifer. All other nutrients were not statistically different between samples. Salinity and pH were higher in the unconfined aquifer, while temperature was lower. Microbial cell counts were similar in the unconfined and confined aquifers (Table 2.1).

2.2.2 Taxonomic and metabolic profiling of groundwater metagenomes

A total of 64,506 and 409,743 sequences with an average read length of 386 and 387 bases were obtained from the unconfined and confined aquifer samples, respectively. Both metagenomic libraries were dominated by bacteria (82% of hits to SEED) (http://metagenomics.theseed.org/) (Overbeek *et al*., 2005) with sequences also matching viruses (9%), archaea (6%), and eukaryota (2%). Proteobacteria represented the highest percentage of matches to the SEED database for both the unconfined and confined aquifers with 18% and 13% of all sequences, respectively (Fig. 2.1A). Within this, the delta/epsilon subdivision contributed to 5% and 7% of the total sequences in the unconfined and confined aquifers, respectively. Viruses (ssDNA) were also major contributors with 3-4% of sequences matching the SEED database (Table S2.1). A total of 278 organisms and 3683 novel sequences could not be assigned to known sequences in the database.

When aquifers were compared using the Statistical Analysis of Metagenomic Profiles (STAMP) software package (Parks and Beiko, 2010), there was an overrepresentation of crenarchaeota, proteobacteria, actinobacteria, chloroflexi,

ssDNA viruses, bacteroidetes/chlorobi group and cyanobacteria in the unconfined aquifer (*q*-value < $1.06e^{-5}$). Conversely, there was an overrepresentation of firmicutes, the fungi/metazoa group and euryarchaeota in the confined aquifer (*q*-value < $1e^{-15}$) (Fig. 2.1B). Similarity percentage (SIMPER) analysis (Clarke, 1993) revealed the main contributors to the dissimilarity between the unconfined and the confined aquifer at phyla level were crenarchaeota and firmicutes, which contributed to 13% and 11% of the dissimilarity respectively (Table S2.2). At finer levels of taxonomic resolution (order level), *Deltaproteobacteria* represented the highest percentage of matches to the SEED database for both unconfined and confined aquifers with 5% and 7% of all sequences, respectively (Fig. 2.2A). STAMP comparisons revealed an overrepresentation of *Rhodospirillales*, *Rhodocyclales*, *Chlorobia* and *Circovirus* occurred in the unconfined aquifer, whereas an overrepresentation of *Deltaproteobacteria* and *Clostridiales* occurred in the confined aquifer (Fig. 2.2B).

In both aquifer samples the core metabolic functions comprising DNA and protein metabolism were most prevalent, while a high level of phosphorus metabolism occurred in the confined aquifer (Table S2.3). Comparisons of the metabolic profiles of the unconfined and confined aquifer using STAMP, revealed an overrepresentation of DNA metabolism in the unconfined aquifer and an overrepresentation of motility and chemotaxis in the confined aquifer (Fig. 2.3A). SIMPER analysis revealed that overall DNA metabolism contributed to 15% of the dissimilarity between the unconfined and confined aquifers, while stress response and motility and chemotaxis contributed approximately 7.5% of the dissimilarity (Table S2.4). Finer levels (subsystem level) of resolution indicated that the unconfined aquifer had an overrepresentation of lactose and galactose uptake and

utilisation, beta-lactamase resistance and DNA replication. The confined aquifer had an overrepresentation of sequences matching sigmaB stress response regulation, flagellum, cobalt-zinc-cadmium resistance, phosphate metabolism and cellulosome degradation (i.e. starch uptake) (Fig. 2.3B).

2.2.3 Comparison of metabolic and taxonomic profiles from other habitats

In order to determine the overall effect the groundwater environment has on the inhabitant microbial assemblages, we compared our groundwater metagenomes to 37 publicly available metagenomes on the MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) pipeline version 2.0 (Meyer *et al*., 2008), covering a wide variety of habitats including other freshwater and low oxygen environments (Table S2.5). The highest metabolism (subsystem) and taxonomy (organism) resolution available was used to create cluster profiles that revealed the unconfined and the confined aquifers were more similar to each other than to any other metagenome (85% and 90% similarity, respectively). When the microbial taxonomy of these samples was compared to metagenomes from other environments, the groundwater samples were most similar to termite gut and cow rumen metagenomes with a cluster node at 75% similarity (Fig. 2.4). When the metabolic potential of these samples was compared to metagenomes from other environments, groundwater samples were most similar to whale fall, phosphorus removing sludge, marine sediment samples and farm soil with a cluster node at 85% similarity (Fig. 2.5).

Table 2.1 Geophysical and microbial enumeration data

Parameter	**Unconfined aquifer (Mean ± SD)** [a]	**Confined aquifer (Mean ± SD)** [a]	***P*-value**
Iron (mg L^{-1})	3.041 ± 0.184	1.232 ± 0.003	0.000 [c]
Sulphur (mg L^{-1})	76.3 ± 4.747	57.5 ± 0.173	0.002 [c]
Ammonia (mg L^{-1})	0.025 ± 0.001	0.023 ± 0.004	0.330
Nitrate (mg L^{-1})	0.012 ± 0.001	0.012 ± 0.011	0.959
Nitrite (mg L^{-1})	0 [b]	0 [b]	-
Phosphorus (mg L^{-1})	0.015 ± 0.001	0.02 ± 0.019	0.718
Total Organic Carbon (mg L^{-1})	2.033 ± 0.208	0.9 ± 0.173	0.002 [c]
Sulphide (mg L^{-1})	0 [b]	0 [b]	-
pH	7.56	7.16	-
Temperature (°C)	16.5	17.54	-
Salinity (ppm)	1.65	1.27	-
Oxygen (mg L^{-1})	0.2	0.26	-
Total Bacterial and Viral Cell Count (cell mL^{-1})	1.15E+05 ± 1.43E+04	1.12E+05 ± 1.08E+04	0.775

[a] Variance is denoted by Standard Deviation.

[b] A value of zero indicates the nutrient is below the detectable limit of the machine. In the case of Nitrite and sulphide this is 0.003 and 0.1mg/L respectively.

[c] Denotes statistically significant values.

2.3 Discussion

2.3.1 Aquifer systems

Aquifer systems are considered to be extreme environments due to a lack of easily accessible organic carbon and low levels of inorganic nutrient input, low oxygen levels and a lack of sunlight (Danielopol *et al*., 2000). Consequently, microbial communities inhabiting these environments consist of microbes adapted to surviving in nutrient poor groundwater environments (Pedersen, 2000). In addition, strong environmental changes driven by anthropogenic influences present a consistent challenge for these communities (Griebler and Lueders, 2009). To determine the effects of anthropogenic influences on groundwater microbes, the microbial ecology of pristine aquifer systems needs to be compared to unconfined aquifers to determine how external factors influence microbial taxonomy and metabolism.

We assessed the chemical properties and the microbial communities within an unconfined aquifer, which has been exposed to external input from a dairy farm, and an adjacent confined aquifer, which has had no external input for approximately 1500 years (Banks *et al*., 2006), to determine the effect of anthropogenic inputs on groundwater ecosystems. Nutrient analysis comparing these two systems showed that the confined aquifer had significantly lower sulphur, iron and total organic carbon (TOC) concentrations than the unconfined aquifer. In groundwater, the amount of suspended microbes is largely dependent on the availability of dissolved organic carbon (DOC) and nutrients (Griebler and Lueders, 2009). Typically phosphorus and iron are limiting factors in groundwater systems (Bennett *et al*., 2001). Those microbes able to increase the bioavailability of such critical nutrients can increase the viability of the native population (Rogers and Bennett, 2004). Flow cytometry counts showed that total bacterial and viral abundances were relatively similar between the

unconfined and confined aquifer with mean values of $1.15 \times 10^5 \pm 1.43 \times 10^4$ and $1.12 \times 10^5 \pm 1.08 \times 10^4$ cells mL $^{-1}$, respectively (Table 2.1). This is consistent with commonly reported microbial cell counts of 10^3 - 10^8 cells mL^{-1} in groundwater regardless of contamination (Pedersen, 1993; 2000; Griebler and Lueders, 2009).

2.3.2 Taxonomic profiling of groundwater

A shift in dominant taxa was observed between the unconfined and the confined aquifer, with fundamentally different communities inhabiting each environment. In the unconfined aquifer there was an overrepresentation of *Rhodospirillales*, *Rhodocyclales*, *Chlorobia* and *Circovirus* (Fig. 2.2). The dominance of these taxa in the unconfined aquifer differs from a recent metagenomic study in which uranium contaminated aquifers were dominated by *Rhodanobacter*-like gammaproteobacterial and *Burkholderia*-like betaproteobacterial species (Hemme *et al*., 2010). However, *Rhodocyclales* are commonly found in wastewater treatment systems (Hesselsoe *et al*., 2009) and are noted for their ability to degrade and transform pollutants such as nitrogen, phosphorus and aromatic compounds (Loy *et al*., 2005). This suggests that the microbial communities in the unconfined aquifer are responding to the influx of nutrients similar to those seen in wastewater. Furthermore, *Chlorobia* are green sulphur bacteria that are typically found in deep anoxic aquatic environments where low light intensity and sulphide concentrations favour their growth (Guerrero *et al*., 2002; Madigan *et al*., 2003). This suggests the increased sulphur concentration in the unconfined aquifer could be responsible for the overrepresented *Chlorobia*. Taken together, these patterns indicate that different types of contamination can drive markedly different community profiles within aquifer system.

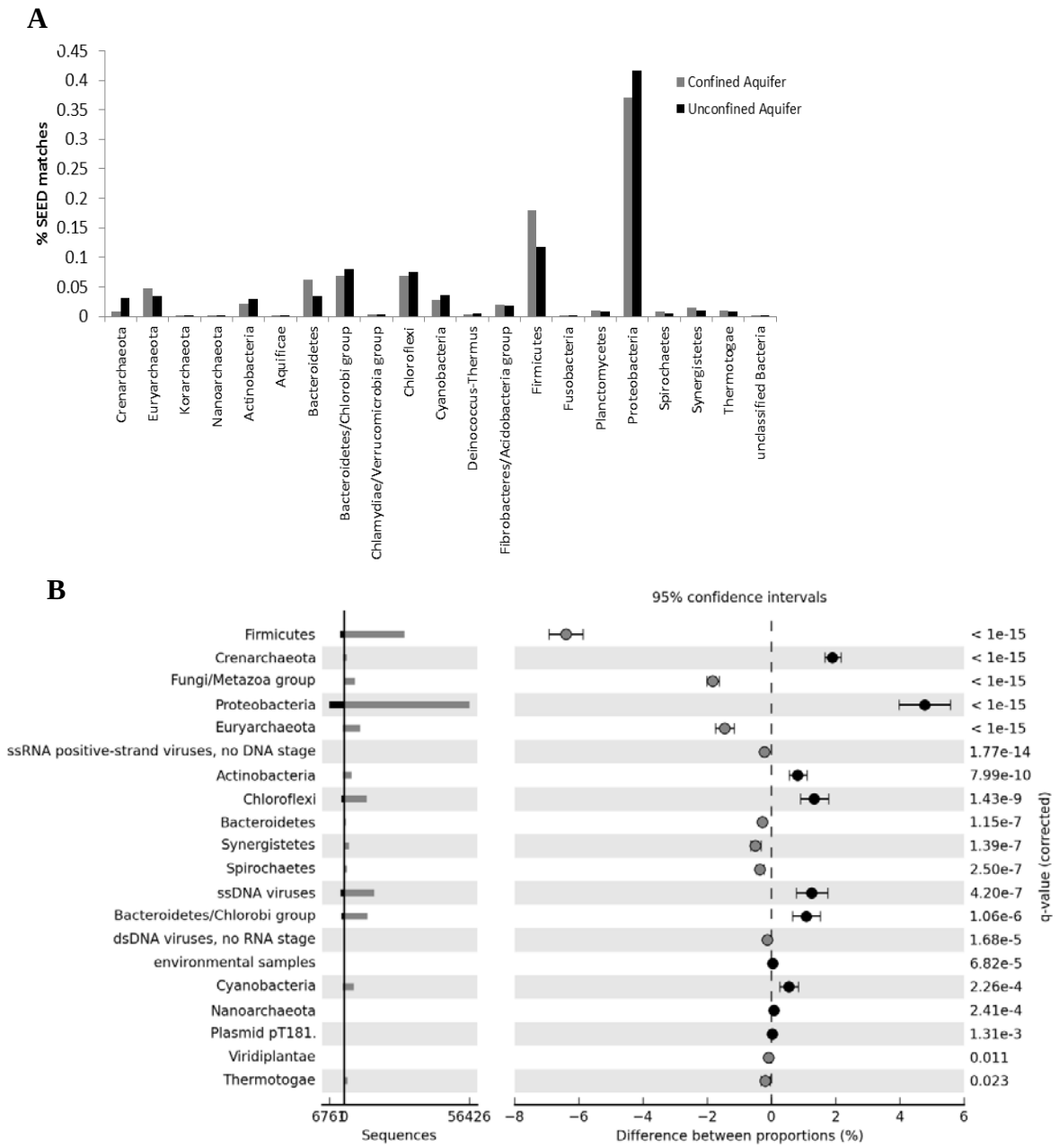

Figure 2.1 Comparison of aquifer taxonomic profiles at phyla level (A) Frequency distribution (relative % of bacterial SEED matches) of bacterial phyla in the unconfined and the confined aquifer. (B) STAMP analysis of taxonomy enriched or depleted between the confined and unconfined aquifers, using approach describes in Parks & Beiko (2010). Groups overrepresented in the unconfined aquifer (black) correspond to positive differences between proportions and groups overrepresented in the confined aquifer (grey) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Storey's FDR approach.

The overrepresentation of circovirus in the unconfined aquifer is also notable, due to its known vertebrate pathogenicity (Rosario *et al*., 2009a). Circoviridae has been linked to a number to livestock related diseases including infections of dairy cattle (Nayar *et al*., 1999) and has previously been found in reclaimed water, suggesting it is resistant to wastewater treatment (Rosario *et al*., 2009b). The occurrence of circoviridae in the unconfined aquifer could indicate contamination from nearby farmland and is consistent with a study by Dinsdale *et al*. (2008a) who found increased numbers of pathogens in a human impacted versus non-human impacted marine environments.

In the confined aquifer there was an overrepresentation of *Deltaproteobacteria* and *Clostridiales* (Fig. 2.2). *Clostridiales* are obligate anaerobes and have the ability to form endospores when growing cells are subjected to nutritional deficiencies (Paredes-Sabja *et al*., 2011). *Clostridiales* have not been widely reported in aquifer systems, however their survival strategies make them well adapted to survive in low nutrient conditions, such as subsurface environments like those observed in the confined aquifer (Leclerc and Moreau, 2002).

2.3.3 Metabolic profiling of groundwater

Generally, the rate of metabolism in subsurface communities is slower in comparison to other aquatic or sediment environments (Swindoll *et al*., 1988). Within groundwater systems, previous studies have shown metabolic rates were higher in a shallow sandy aquifer compared to a confined clayey aquifer (Chapelle and Lovley, 1990). The authors suggested this lower metabolism could be due to the reduced interconnectivity, and thus, a reduction in microbial and nutrient mobility. The core metabolic function in each of our aquifer systems was DNA metabolism; however an

overrepresentation of DNA replication was seen in the unconfined aquifer compared to the confined (Fig. 2.3). This indicates that the reduced nutrient levels in the confined aquifer may have led to reduced reproduction.

When nutrient levels are low, it is advantageous for microbes to attach themselves to sediment particles, detritus, rock surfaces and biofilms (Griebler and Lueders, 2009). This attachment mode is successful as nutrient availability is higher at surfaces (Hall-Stoodley *et al*., 2004). Thus, microbes dominating groundwater systems are more commonly found attached to surfaces than in suspension (Griebler and Lueders, 2009). Repulsive forces of the substratum require microbial cells to produce flagella for the early stages of attachment (Donlan, 2002). Overrepresentation of flagella in the confined aquifer community (Fig. 2.3) could be indicative of a greater need to attach to surfaces in the low nutrient confined aquifer.

Our data also indicate that β-lactamase genes were overrepresented in the unconfined aquifer (Fig. 2.3). This antibiotic resistance gene is widely seen in Gram-negative bacteria and has been shown to be a product of the extensive use of β-lactams in dairy farms to prevent bacterial infections (Berghash *et al*., 1983; Gianneechini *et al*., 2002; Sawant *et al*., 2005; Liebana *et al*., 2006). Within live-stock, the majority of antibiotics are excreted unchanged by the animal, where they subsequently enter water sources via leaching and run-off (Zhang *et al*., 2009). This has caused concern about the potential impacts that antibacterial resistance in waterways can have on humans and animal health (Kemper, 2008). The overrepresentation of β-lactamase in the unconfined aquifer suggests that external input, potentially in the form of farm affected input, may introduce new cellular processes that would not normally be required by endemic groundwater microbes. This is consistent with a study that

investigated the use of antibiotics in farm animals and illustrated that antibiotic resistance can be spread into the surrounding environment through the use of antimicrobial drugs (Ghosh and LaPara, 2007). Further, microbes able to utilize lactose have previously been linked to dairy farms (Klijn *et al*., 1995) and thus, the overrepresentation of lactose and glucose utilization found in the unconfined aquifer (Fig. 2.3) could be linked to external input from the overlaying dairy farms.

2.3.4 Comparison to other microbial communities

To determine how the unique features of the groundwater environment influence the structure of microbial communities, we compared the metagenomes from our aquifer systems to metagenomes from different environments (Table S2.5). The unconfined and confined aquifer metagenomes were more similar to each other than to any other community, both in terms of taxonomy and metabolism (Fig. 2.4 and 2.5). This suggests the features of subterranean aquatic environments, including low oxygen concentrations, coupled with a lack of sunlight and low external inputs of nutrients have led to a unique niche for microbial communities to evolve. In a recent study, four sediment metagenomes from a naturally occurring salinity gradient were compared and it was found that despite differences in salinity and nutrient levels, these four samples clustered more closely to each other and other sediment samples, than to other similar hypersaline environments (Jeffries *et al*., 2011a). It was found that the substrate type, i.e. sediment or water, rather than salinity drove the similarity. Willner *et al*. (2009) also found that microbiomes and viromes have distinct sequence-based signatures which are driven by environmental selection. This is further supported by Dinsdale *et al*. (2008b), who compared metagenomic sequences from 45 distinct microbiomes and 42 distinct viromes to show there was a strong discriminatory profile across different environments. Our data similarity suggest that

the unique features of the subterranean aquatic environment act to structure microbial assemblages that retain a high level of similarity between different aquifers.

The taxonomy of the aquifer metagenomes were most similar to cow rumen and termite gut metagenomes (Fig. 2.4). A common feature among these environments is the incidence of anaerobic fungi which is overrepresented in the confined aquifer (Fry *et al*., 1997; Ramšak *et al*., 2000; Ekendahl *et al*., 2003; Warnecke *et al*., 2007). A primary role of anaerobic fungi in gut systems is the large scale break-down of plant material, including cellulose (Ramšak *et al*., 2000; Warnecke *et al*., 2007). The breakdown of cellulose in groundwater is also known to occur in shallow aquifers (Vreeland *et al*., 1998) which along with the overrepresentation in cellulosome genes in the confined aquifer (Fig. 2.3), suggests that cellulose is present and possibly an important food source for the overrepresented fungi/metazoa group (Fig. 2.1). Furthermore, the cellulosome gene is similarly represented in the groundwater, termite gut and cow rumen, suggesting cellulose is a major factor linking the three environmental metagenomes.

The metabolism of the aquifer metagenomes were most similar to other sediment metagenomes (85% similar) rather than freshwater environments (80% similar) (Fig. 2.5). Common features to groundwater and sediment environments are low oxygen concentrations, a lack of sunlight and large surfaces for biofilm formation (Griebler and Lueders, 2009). As previously discussed, due to low nutrient levels in groundwater environments, a common survival strategy is for the microbes to attach to sediment particles or form biofilms (Hall-Stoodley *et al*., 2004; Griebler and Lueders, 2009). This suggests, the attachment mode of life coupled with the low

oxygen concentrations and a lack of sunlight, are the main factors driving the similarity between these metagenomes.

2.3.5 Caveats

Due to the low microbial biomass in groundwater systems, we used multiple displacement amplification (MDA) prior to 454 pyrosequencing. This method has been used widely to amplify DNA prior to sequencing (Binga *et al*., 2008; Dinsdale *et al*., 2008a; Neufeld *et al*., 2008; Palenik *et al*., 2009), but its suitability for use in quantitative metagenomic analysis has been debated (Yilmaz *et al*., 2010) because of the GC bias introduced. However, in our study, as GenomiPhi was used on both aquifer samples compared here, any bias in the process is applied to both aquifers. Furthermore, we are concerned with differences between aquifer groups rather than absolute changes in particular genes. Edwards *et al*. (2006) used GenomiPhi to amplify microbial DNA from a Soudan Mine and found that the whole genome amplification bias was minimal and was found preferentially towards the ends of linear DNA. The authors concluded that as these biases were applied equally to both libraries, this bias would have been negated during the comparative study when assessing differences in the community structure (Edwards *et al*., 2006).

There is a possibility that the clustering of our samples may be due to the way in which the samples were collected, sequenced and analysed, which may be different to the metagenomes from other environments. However, there is no evidence of clustering based on collection, DNA extraction, MDA or sequencing protocols (Fig. 2.4 and 2.5), and thus a technical bias is not evident.

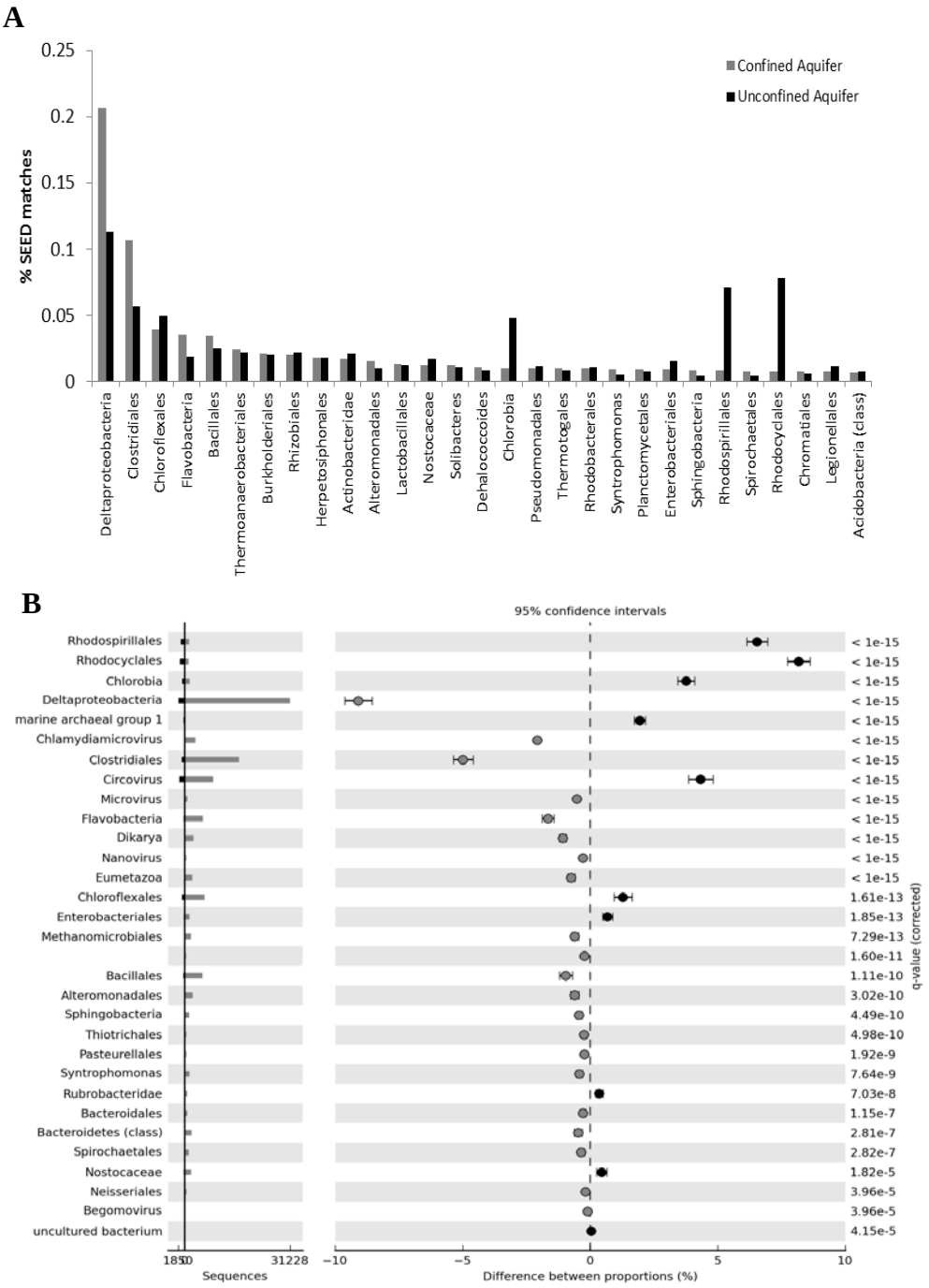

Figure 2.2 Comparison of aquifer taxonomic profiles at order level taxonomy

(A) Frequency distribution (relative % of bacterial SEED matches) of taxonomy in the unconfined and the confined aquifer. (B) STAMP analysis of taxonomy enriched or depleted between the confined and unconfined aquifers. Groups overrepresented in the unconfined aquifer (black) correspond to positive differences between proportions and groups overrepresented in the confined aquifer (grey) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Storey's FDR approach.

2.4 Conclusion

Our data indicates that aquifer ecosystems host unique microbial assemblages that have different phylogenetic and metabolic properties to other environments. We suggest this pattern is driven by the unique physio-chemical properties of subterranean aquatic environments, and that groundwater ecosystems represent a specific microbial niche. Our data also revealed that the unconfined aquifer examined in this study has significantly different features to the more pristine confined aquifer, which in some cases appear to have been influenced by external input from a surrounding dairy farm. Increased nutrient concentrations, the overrepresentation of DNA replication as well as lactose and galactose utilization and β-lactamase genes are all consistent with inputs of nutrients and contaminants from dairy farm practises. Preservation of groundwater is of increasing importance due to its use as potable water sources and as water sources for global industrial and agricultural production. This study provides important insights and suggests further investigation into the differences between unconfined and confined aquifers. Further to this, a study of the subterranean dispersal of agricultural contaminants is needed in order to fully determine the effects of anthropogenic processes on groundwater.

2.5 Experimental Procedures

2.5.1 Site selection

Samples were collected from two depths in the Ashbourne aquifer system, situated within the Finniss River Catchment, South Australia (35°18'S 138°46'E) in June 2010. The Ashbourne aquifer system is two aquifer ecosystems with separate recharge processes that have distinct water sources. The confined aquifer has been isolated from external input for approximately 1500 years (Banks *et al*., 2006), and thus provides a baseline for which the unconfined aquifer can be compared.

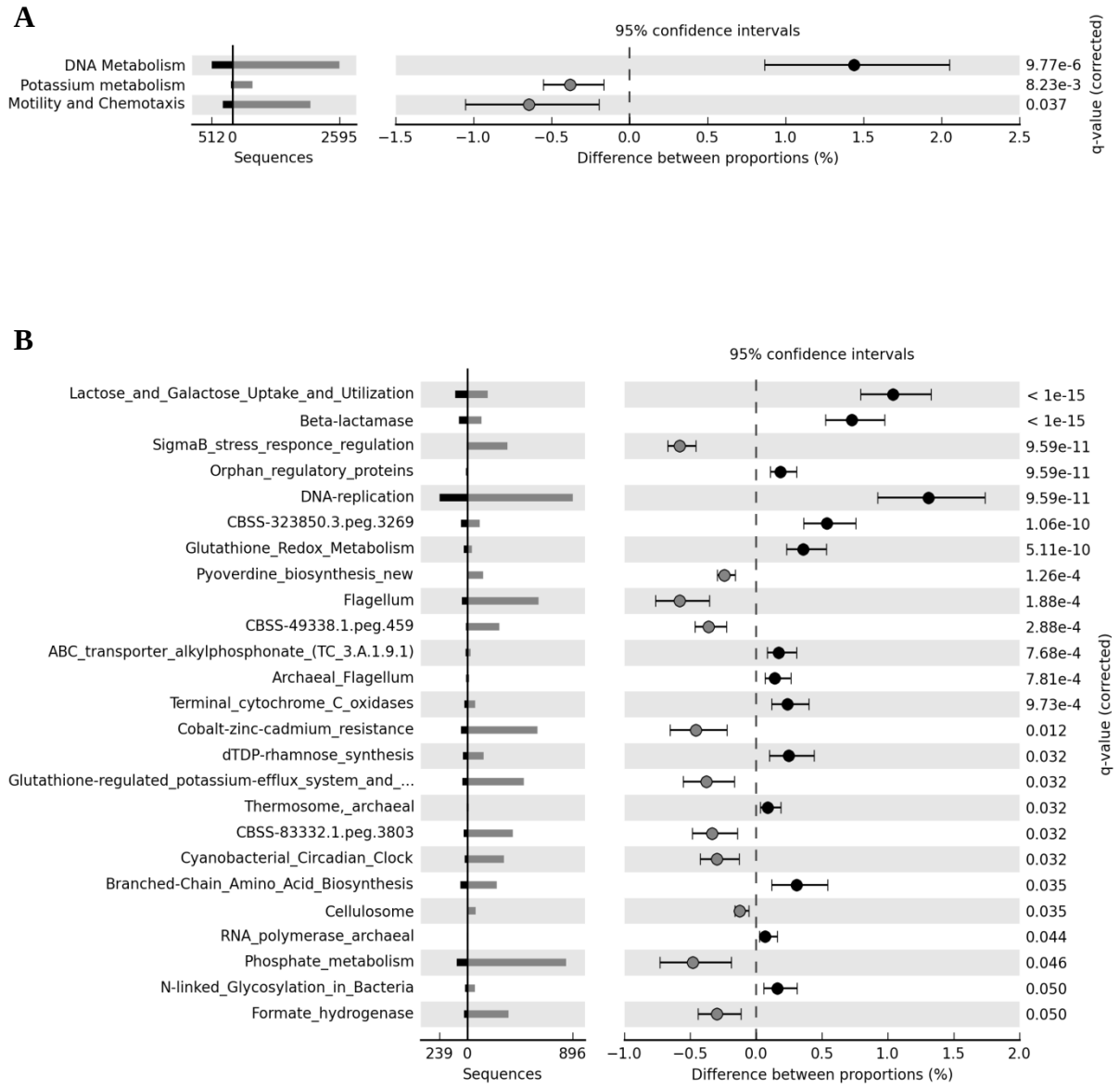

Figure 2.3 Comparison of aquifer metabolism profiles (A) STAMP analysis of hierarchy 1 enriched or depleted between the confined and unconfined aquifers. Groups overrepresented in the unconfined aquifer (black) correspond to positive differences between proportions and groups overrepresented in the confined aquifer (grey) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Storey's FDR approach. (B) STAMP analysis of subsystems enriched or depleted between the confined and unconfined aquifers. Groups overrepresented in the unconfined aquifer (black) correspond to positive differences between proportions and groups overrepresented in the confined aquifer (grey) correspond to negative differences between proportions.

2.5.2 Sampling Groundwater

Unconfined and confined aquifer samples were collected from a nested set of piezometers. Each piezometer consisted of a 10 mm diameter PVC casing, with slotted PVC screens that provide discrete sampling points at specific depths. The unconfined aquifer was sampled from a piezometer at 13-19 m and the confined aquifer at 79-84 m. To ensure that only aquifer water was sampled, bores were purged by pumping out 3 bore volumes using a 12 V, 36 m monsoon pump (EnviroEquip, Inc.) prior to sampling. Based on microbial abundances at each depth determined previously using flow cytometry, 20 L and 200 L of water was collected from the unconfined and confined aquifers respectively, to ensure sufficient biomass for microbial DNA recovery.

From each sampling location, triplicate 600 mL water samples for inorganic and organic chemistry analysis were collected and stored on ice. Nutrient analysis for ammonia, nitrite, nitrate, and filterable reactive phosphorus were conducted using a flow injection analyser. TOC was analysed using OI analytical 1010 & 1030 low level TOC analysers, iron and sulphur were determined by the ICP-006 and ICP-004 elemental analysis using an ICP-mass spectrometer, and sulphide (S^{2-}) concentrations were determined using the colorimetric method (APHA 1995). All analysis was conducted at the Australian Water Quality Centre (Adelaide). For enumeration of microbes at each site, triplicate 1 mL samples were fixed with gluteraldehyde (2% final concentration), quick frozen in liquid nitrogen and stored at -80°C prior to flow cytometric analysis (Brussaard, 2004). Physical parameters, including temperature, salinity, pH, and oxygen concentration, were recorded at each sampling point with the use of a MS5 water quality sonde (Hach Hydrolab®).

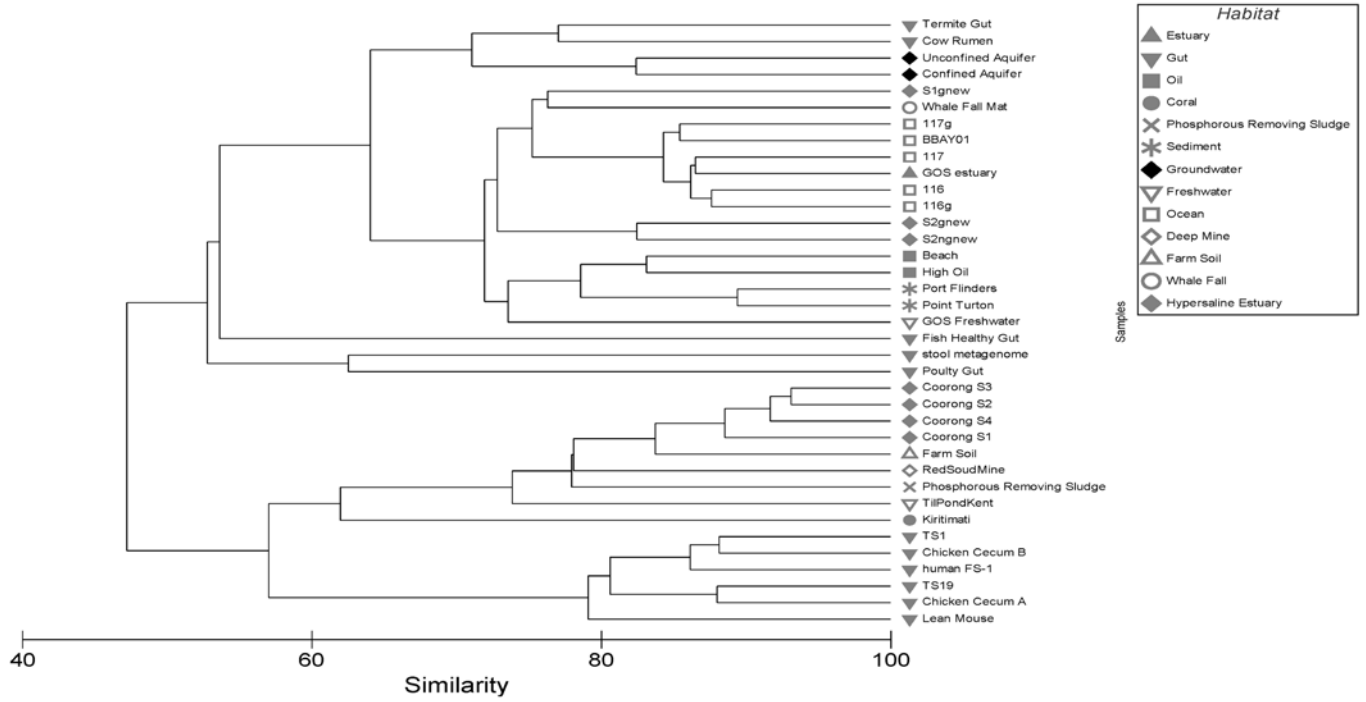

Figure 2.4 Comparison of aquifer taxonomic profiles along with publicly available profiles available on the MG-RAST database. Cluster plot is derived from a Bray-Curtis similarity matrix calculated from the square-root transformed abundance of DNA fragments matching genome level taxonomy in the SEED database (BLASTX E-value <0.001). Details of metagenomes are in Table S2.5.

2.5.3 Microbial enumeration

Bacteria and viruses were enumerated using a FACSCanto flow cytometer (Becton-Dickson). Prior to analysis, triplicate samples were quick thawed and diluted 1:10 with 0.2 μm filtered TE buffer (10 mM Tris, 1 mM EDTA pH 7.5). Samples were then stained with SYBR-I Green solution (1:20000 dilution; Molecular Probes, Eugene, OR) and incubated in the dark for 10 min at 80°C (Brussaard, 2004). As an internal size standard fluorescent 1 μm diameter beads (Molecular Probes, Eugene, OR) were added to each sample at a final concentration of approximately 10^5 beads mL^{-1} (Gasol and Del Giorgio, 2000). Forward scatter (FSC), side scatter (SSC) and green (SYBR®Green-I) fluorescence were acquired for each sample. WinMDI 2.9 (© Joseph Trotter) software was used to identify and enumerate microbes according to variations in green fluorescence and side scatter (Marie *et al*., 1997; 1999; Brussaard, 2004).

2.5.4 Sample filtration, microbial community DNA extraction and sequencing

Following collection, samples for metagenomic analysis were filtered through 5 μm membranes to remove sediment particles before being concentrated by 2000-fold using a 100 kDa tangential flow filtration (TFF) filter (MilliporeTM). Microbial community DNA was extracted using a bead beating and chemical lysis extraction protocol (PowerWater® DNA Isolation Kit; MoBio laboratories, Inc.). Due to the low microbial biomass in the aquifer samples, DNA was then amplified using the multiple strand displacement Phi29 DNA polymerase (GenomiPhi V2 Kit; GE Healthcare Life Sciences, Inc.) and cleaned up using a PCR clean-up kit (UltraClean® PCR Clean-Up Kit; MoBio laboratories, Inc.). DNA quality and concentration were determined by 1.5% TBE agarose gel electrophoresis (Bioline) and a Qubit fluorometer (Quant-iT™ dsDNA HS Assay Kit; Invitrogen Inc.).

Approximately 500 ng of high molecular weight DNA was then sequenced by the Ramaciotti Centre for Gene Function Analysis, Sydney, Australia. Sequencing was conducted on the GS-FLX pyrosequencing platform using Titanium series reagents (Roche).

2.5.5 Data analysis

To determine if the nutrient data was statistically different between the unconfined and the confined aquifer, *P*-values were determined by an Independent t-test. All analysis was performed using PASW version 18 statistical software.

Unassembled DNA sequences were annotated with the MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) pipeline version 2.0 (Meyer *et al*., 2008). BLASTX was used with a minimum alignment length of 50 bp and an E-value cut-off of $E<1e^{-5}$ as described by Dinsdale *et al*. (2008b). Taxonomic profiles were generated using the normalized abundance of sequence matches to the SEED database (Overbeek *et al*., 2005), while the normalized abundance of sequence matches to a given subsystem were used to generate metabolic profiles.

To determine statistically significant differences between the two aquifer samples, the Statistical Analysis of Metagenomic Profiles (STAMP) software package was used (Parks and Beiko, 2010). First, a table of the frequency of hits to each individual taxa or subsystem for each metagenome was generated, which had been normalised by dividing by the total number of hits to remove bias in difference in read lengths and sequencing effort. An E-value cut-off of $E<1e^{-5}$ was used to identify hits. The highest level of resolution available on MG-RAST was used for metabolism (subsystem) and taxonomy (genome). *P*-values were calculated in STAMP using the two sided Fisher's Exact test (Fisher, 1958), while the confidence intervals were

calculated using the Newcombe-Wilson method (Newcombe, 1998). False discovery rate was corrected for using the Storey's FDR method (Storey and Tibshirani, 2003).

We next compared the metagenomes of our groundwater samples to 37 publicly available metagenomes from a variety of environments on MG-RAST (Table S2.5), to statistically investigate the similarities between the two groundwater samples as well as other environments. Heatmaps were generated and normalized, as described above; however, as groundwater samples were compared to datasets with a variety of different read lengths, a lower E-value cut-off of E<0.001 was used. Statistical analyses were conducted on square-root transformed data using the statistical package Primer 6 for Windows (Version 6.1.6, Primer-E Ltd. Plymouth) (Clarke and Gorley, 2006). Metagenomes were then analysed using hierarchial agglomerative clustering (CLUSTER) (Clarke, 1993) analyses of the Bray-Curtis similarities. The main taxa or subsystems contributing to the differences were identified using similarity percentage (SIMPER) analysis (Clarke, 1993).

2.6 Acknowledgments

The authors gratefully acknowledge Eugene Ng from the Flow Cytometry Unit of the Flinders University Medical Centre for providing technical support during the flow cytometry work. Funding was provided by ARC linkage grant LP0776478. Renee Smith is the recipient of a Flinders University Research Scholarship (FURS).

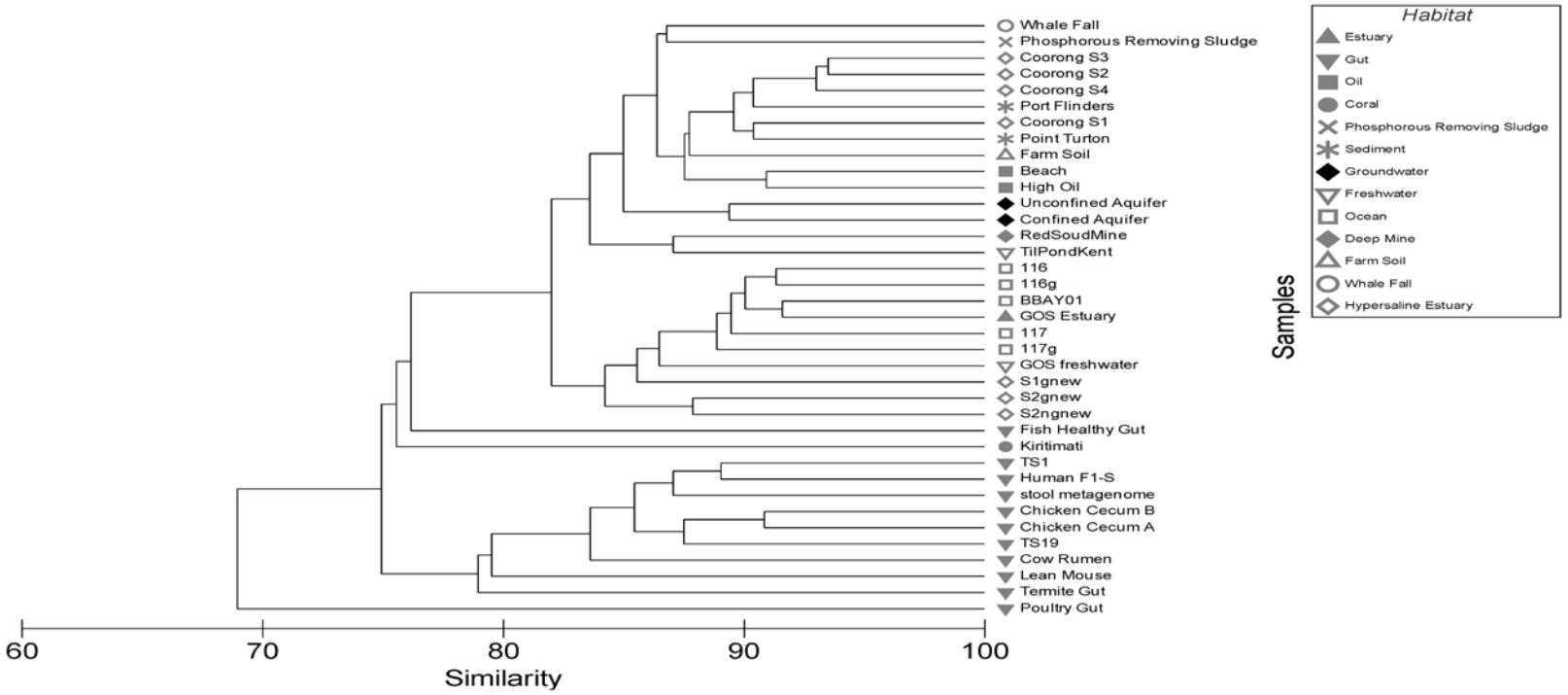

Figure 2.5 Comparison of aquifer metabolic profiles along with publicly available profiles available on the MG-RAST database. Cluster plot is derived from a Bray-Curtis similarity matrix calculated from the square-root transformed abundance of DNA fragments matching subsystems in the SEED database (BLASTX E-value <0.001). Details of metagenomes are in Table S2.5.

Table S2.1 Relative proportion of matches to the SEED database taxonomic hierarchy.

Domain	MG-RAST Level 2 (Phyla)	MG-RAST Level 3	Confined aquifer	Unconfined aquifer
Bacteria	Proteobacteria	Delta/epsilon subdivision	0.2109	0.1182
Bacteria	Firmicutes	Clostridia	0.1307	0.0784
Bacteria	Proteobacteria	Gammaproteobacteria	0.0746	0.0786
Bacteria	Chloroflexi	Chloroflexi (class)	0.0573	0.0675
Viruses	ssDNA viruses	Circoviridae	0.057	0.0904
Bacteria	Firmicutes	Bacilli	0.0475	0.0378
Bacteria	Proteobacteria	Alphaproteobacteria	0.0434	0.1096
Bacteria	Proteobacteria	Betaproteobacteria	0.0393	0.1086
Archaea	Euryarchaeota	Methanomicrobia	0.0279	0.0188
Viruses	ssDNA viruses	Microviridae	0.0269	0.0003
Bacteria	Actinobacteria	Actinobacteria	0.022	0.0295
Bacteria	Fibrobacteres/ Acidobacteria group	Acidobacteria	0.0191	0.0184
Bacteria	Bacteroidetes	Bacteroidetes (class)	0.0184	0.0106
Eukaryota	Fungi/Metazoa group	Fungi	0.016	0.0066
Bacteria	Synergistetes	Syntrophomonadaceae	0.014	0.0092
Bacteria	Cyanobacteria	Nostocales	0.0127	0.0171
Eukaryota	Fungi/Metazoa group	Metazoa	0.0124	0.0077
Bacteria	Bacteroidetes/Chlorobi group	Chlorobi	0.0118	0.0498
Bacteria	Chloroflexi	Dehaloccoidetes	0.011	0.0084
Bacteria	Planctomycetes	Planctomycetacia	0.0091	0.0079
Bacteria	Cyanobacteria	Chroococcales	0.0086	0.0124
Bacteria	Spirochaetes	Spirochaetes (class)	0.0079	0.0046
Bacteria	Thermotogae	Thermotogae (class)	0.0079	0.0073
Archaea	Crenarchaeota	Thermoprotei	0.0075	0.0318
Archaea	Euryarchaeota	Thermococci	0.0041	0.0039
Bacteria	Deinococcus-Thermus	Deinococci	0.004	0.0049
Archaea	Euryarchaeota	Methanobacteria	0.0039	0.0027
Viruses	ssDNA viruses	Geminiviridae	0.0034	0.0015
Archaea	Euryarchaeota	Methanococci	0.0033	0.0021
Archaea	Euryarchaeota	Archaeoglobi	0.0031	0.0018
Viruses	ssDNA viruses	Nanoviridae	0.0031	0.0001
Bacteria	Cyanobacteria	Gloeobacteria	0.003	0.0035
Viruses	Bacteriophage phBC6A51		0.003	0.0007
Bacteria	Cyanobacteria	Oscillatoriales	0.0029	0.0027
Bacteria	Chlamydiae/ Verrucomicrobia group	Chlamydiae	0.0027	0.0021

Bacteria	Proteobacteria	Unclassified Proteobacteria	0.0027	0.0022
Bacteria	Aquificae	Aquificae (class)	0.0024	0.0023
Viruses	ssRNA positive-strand viruses, no DNA stage	Sclerophthora macrospora virus A.	0.0024	0
Archaea	Euryarchaeota	Halobacteria	0.0023	0.0027
Eukaryota	Viridiplantae	Streptophyta	0.0021	0.0012
Archaea	Korarchaeota	Candidatus Korarcheaum	0.0018	0.0014
Bacteria	Fusobacteria	Fusobacteria (class)	0.0018	0.0013
Archaea	Euryarchaeota	Thermoplasmata	0.0017	0.0014
Bacteria	Unclassified bacteria	Candidate division TG1	0.0017	0.0015
Bacteria	Chlamydiae/ Verrucomicrobia group	Verrucomicrobia	0.0012	0.0013
Viruses	dsDNA viruses, no RNA stage	Caudovirales	0.0011	0.0003
Archaea	Euryarchaeota	Methanopyri	0.001	0.001
Bacteria	Cyanobacteria	Prochlorales	0.001	0.0007
Bacteria	Firmicutes	Mollicutes	0.0008	0.0006
Viruses	dsDNA viruses, no RNA stage	Poxviridae	0.0007	0.0003

Top 50 hits were generated by BLASTing sequences to the SEED database with a minimum alignment length of 50 bp and an E-value cut-off of $1e^{-5}$.

Relative representation in the metagenome was calculated by dividing the number of hit to each category by the total number of hits to all categories.

Table S2.2 Contribution of phyla level taxonomy to the dissimilarity of confined and unconfined aquifer metagenomes.

Species	Avg. Abundance		Contribution %	Cumulative %
	Unconfined aquifer	Confined aquifer		
Crenarchaeota	**0.18**	0.09	12.94	12.94
Firmicutes	0.34	**0.42**	11.01	23.94
Bacteriodetes	0.19	**0.25**	9.53	33.47
Fungi/Metazoa group	0.13	**0.19**	8.09	41.56
ssRNA positive-strand viruses, no DNA stage	0	**0.05**	7.02	48.58
Proteobacteria	**0.65**	0.61	5.86	54.43
Euryarchaeota	0.19	**0.22**	4.52	58.95

Percentage differences calculated using SIMPER analysis.

Table S2.3 Relative proportion of matches to a given subsystem hierarchy 1.

Subsystem Hierarchy 1	Confined aquifer	Unconfined aquifer
Phosphorus metabolism	0.0173	0.0123
DNA metabolism	0.0164	0.0296
Protein metabolism	0.0157	0.0173
Motility and chemotaxis	0.0149	0.0113
Regulation and cell signalling	0.0132	0.011
Clustering-based subsystems	0.0129	0.0138
Stress response	0.0119	0.0015
Motility and chemotaxis	0.0017	0.0087
Respiration	0.0114	0.0092
Virulence	0.0113	0.0075
Unclassified	0.0107	0.0087
Motility and chemotaxis	0.0104	0.0105
DNA metabolism	0.0102	0.0111
Respiration	0.0098	0.006
Cell wall and capsule	0.0097	0.0086
Potassium metabolism	0.0094	0.0072
Stress response	0.0084	0.012
Membrane transport	0.0082	0.0048
DNA metabolism	0.0079	0.0065
Virulence	0.0078	0.0082
Nucleosides and Nucleotides	0.0077	0.0093
Unclassified	0.0075	0.0047
Cofactors, vitamins, prosthetic groups, pigments	0.0072	0.0064
Carbohydrates	0.0071	0.0064
Amino acids and derivatives	0.0071	0.0085
Carbohydrates	0.007	0.0075
Cell division0.0068 and cell cycle	0.0068	0.0075
Miscellaneous	0.0065	0.0072
Respiration	0.0064	0.0044
Clustering-based subsystems	0.0063	0.0049
Clustering-based subsystems	0.0062	0.0055
Protein metabolism	0.0062	0.0092
Cell division and cell cycle	0.0061	0.0064
Carbohydrates	0.006	0.0051
Cell division and cell cycle	0.0059	0.0033
Clustering-based subsystems	0.0056	0.0078
Clustering-based subsystems	0.0055	0.0045
Respiration	0.0055	0.0048
Protein metabolism	0.0054	0.0058
Cofactors, vitamins, prosthetic groups, pigments	0.0052	0.0045
Clustering-based subsystems	0.0052	0.0068
Clustering-based subsystems	0.005	0.0019
Protein metabolism	0.005	0.0055
Nucleosides and nucleotides	0.005	0.0057

Carbohydrates	0.0049	0.0029
Stress response	0.0049	0.0033
Amino acids and derivatives	0.0049	0.0036
Virulence	0.0049	0.0044
Clustering-based subsystems	0.0049	0.0061
Virulence	0.0048	0.0047

Top 50 hits were generated by BLASTing sequences to the MG-RAST subsystem database with a minimum alignment length of 50 bp and an E-value cut-off of $1e^{-5}$.

Relative representation in the metagenome was calculated by dividing the number of hit to each category by the total number of hits to all categories.

Table S2.4 Contribution of metabolic hierarchical 1 system to the dissimilarity of confined and unconfined aquifer metagenomes.

	Avg. Abundance			
Metabolic Processes	**Unconfined aquifer**	**Confined aquifer**	**Contribution %**	**Cumulative %**
DNA metabolism	0.26	0.22	14.99	14.99
Stress response	0.18	0.2	7.85	22.85
Motility and chemotaxis	0.18	0.2	7.67	30.51

Percentage differences calculated using SIMPER analysis.

Table S2.5 Summary of publicly available metagenomes used in this study.

MG-RAST ID	Description/Reference	MG-RAST ID	Description/Reference
4453064.3	Unconfined aquifer	4444843.3	Poultry Gut
4453083.3	Confined aquifer	4441695.3	Fish healthy gut (Angly *et al*., 2009)
4440984.3	Coorong sediment 1	4440283.3	Chicken cecum A (Qu *et al*., 2008)
4441020.3	Coorong sediment 2	4440284.3	Chicken cecum B (Qu *et al*., 2008)
4441021.3	Coorong sediment 3	4440452.7	TS1 (human gut) (Turnbaugh *et al*., 2009)
4441022.3	Coorong sediment 4	4440610.3	TS19 (human gut) (Turnbaugh *et al*., 2009)
4446406.3	Coorong water 1	4440939.3	Human FS-1 (human gut) (Kurokawa *et al*., 2007)
4446412.3	Coorong water 2	4440463.3	Lean mouse (gut) (Turnbaugh *et al*., 2006)
4446411.3	Coorong water 3	4444130.3	Stool
4446457.3	East Australian Current 1 (Seymour *et al*., 2012)	4441656.4	Whalefall mat (Tringe *et al*., 2005)
4446409.3	East Australian Current 2 (Seymour *et al*., 2012)	4440281.3	Soudan mine (Edwards *et al*., 2006)
4446407.3	East Australian Current 3 (Seymour *et al*., 2012)	4441091.3	Farm soil (Edwards *et al*., 2006)
4446410.3	East Australian Current 4 (Seymour *et al*., 2012)	4443688.3	Botany Bay (marine)
4446341.3	Marine sediment 1	4440041.3	Kiritimati (marine) (Dinsdale *et al*., 2008a)
4446342.3	Marine sediment 2	4441584.3	GS012 (estuary) (Rusch *et al*., 2007)
4453072.3	Oil contaminated soil 1	4441590.3	GS020 (freshwater) (Rusch *et al*., 2007)
4453082.3	Oil contaminated soil 2	4440440.3	Aquaculture pond (Dinsdale *et al*., 2008b)
4442701.3	Termite gut (Warnecke *et al*., 2007)	4441092.3	Phosphorus removing sludge
4441682.3	Cow Rumen (Brulc *et al*., 2009)		

Chapter 3

Confined aquifers as viral reservoirs

Submitted as:

Smith RJ, Jeffries TC, Roudnew B, Seymour JR, Fitch AJ, Speck PG, Newton K, Brown MH, Mitchell JG (2012) Confined aquifers as viral reservoirs. *Environmental Microbiology Reports* (In Review).

3.0 Summary

Potentially pathogenic viruses within freshwater reserves represent a global health risk. However, knowledge about their diversity and abundance in deep groundwater reserves is currently limited. We found that the viral community inhabiting a deep confined aquifer in South Australia was more similar to reclaimed water communities than to the viral communities in the overlying unconfined aquifer community. This similarity was driven by high relative occurrence of the ssDNA viral groups *Circoviridae*, *Geminiviridae*, *Inoviridae* and *Microviridae*, which include many known plant and animal pathogens. These groups were present in 1500 year-old water situated 80 m below the surface, which suggests the potential for long-term survival and spread of potentially pathogenic viruses in deep, confined groundwater. Obtaining a broader understanding of potentially pathogenic viral communities within aquifers is particularly important given the ability of viruses to spread within groundwater ecosystems.

3.1 Introduction

Confined aquifers typically lie deep below the surface and are permanently, or semi-permanently, separated from other groundwater by low permeability geologic formations, which provide barriers to flow (Hamblin and Christiansen, 2004; Borchardt *et al*., 2007). These barriers are thought to protect the underlying groundwater from the overlying environment, and thus prevent the spread of contaminants into the freshwater reserves (Nolan *et al*., 1997). However, vertical fractures can lead to the formation of pathways for water movement, allowing for the introduction of surface contaminants, including microbial pathogens (Eaton *et al*., 2007). Among microbial pathogens, enteric viruses have substantial potential for spread into deep aquifers due to their small, 27 – 75 nm, size (Borchardt *et al*., 2007).

Human pathogens within freshwater reserves are a global health risk (Toze, 1999; Abbaszadegan *et al*., 2003). The persistence and viability of pathogenic viruses can vary widely based on the surrounding environment (Ouellette *et al*., 2010). Some reports indicate that viruses can remain in an infectious state within deep groundwater for years, but that they become unviable in surface waters after only a few days (Borchardt *et al*., 2007; Nazir *et al*., 2010). Enhanced virus viability and longevity within deep groundwater may be related to the lower temperatures and a lack of sunlight in this habitat (Yates *et al*., 1985; Diels, 2005), as well as the attachment of viruses to surfaces (Sim and Chrysikopoulos, 2000). This longevity, along with their 20 – 350 nm size, means that viruses have higher potential dispersal levels within groundwater systems than bacteria (Scheuerman *et al*., 1987; Diels, 2005). The distance viruses can spread and the time they can remain in groundwater is poorly understood and will depend on the biological and physical conditions of specific groundwater systems. One of the first steps in understanding the potential for

dispersal is identifying the occurrence of deep water pathogenic viruses. Therefore, it is important to determine the identity of viruses within groundwater ecosystems.

A recent metagenomic study of an aquifer system revealed a relatively high proportion of viral sequences, 9% (Smith *et al*., 2011), when compared to other aquatic environments, 0.1-1% (Edwards and Rohwer, 2005; Jeffries *et al*., 2011a). Therefore, we sought to construct a viral community profile from the viral sequences in the unconfined and confined aquifer metagenomes, including the discrimination of any potential human pathogens. This data was compared to metagenomes from a number of other marine and freshwater environments.

3.2 Results and Discussion

Groundwater samples were collected from the confined and unconfined Ashbourne aquifer systems, South Australia (35°18'S 138°46'E) in June 2010. The unconfined aquifer is exposed to overlying input, while the confined aquifer lies at 40 m, below a 15 m thick confining layer, and has been isolated from external input for approximately 1500 years (Banks *et al*., 2006). Separate recharge processes have led to distinct water sources that differ between the confined and unconfined aquifers (Banks *et al*., 2006; Smith *et al*., 2011). Metagenomes were sequenced using the GS-FLX pyrosequencing platform using Titanium reagents (Roche). The resulting 409,743 and 64,506 sequences from the confined and unconfined aquifers, respectively, were compared to the Viral Proteins database in the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) pipeline (Seshadri *et al*., 2007). BLASTX and an $E < 1 \times 10^{-5}$ was used to identify hits.

Table 3.1 Summary of publicly available metagenomes used in this study.

Database	Description	Reference
MG-RAST	Unconfined Aquifer	(Smith *et al*., 2011)
MG-RAST	Confined Aquifer	(Smith *et al*., 2011)
MG-RAST	Danish Wastewater Treatment Plant	(Albertsen *et al*., 2012)
MG-RAST	Botany Bay	(Burke *et al*., 2011)
CAMERA	Viral Metagenome from reclaimed water	(Rosario *et al*., 2009b)
CAMERA	Chesapeake Bay Virioplankton Metagenome	(Bench *et al*., 2007)
CAMERA	Viral Metagenome from the Freshwater Lake Limnopolar	(López-Bueno *et al*., 2009)
CAMERA	Viral Metagenomes from Terrestrial Hot Springs	(Schoenfeld *et al*., 2008)
CAMERA	Viral Stromatolite Metagenome	(Desnues *et al*., 2008)
CAMERA	Wastewater	(Sanapareddy *et al*., 2008)

The majority of viral sequences within our confined and unconfined aquifer metagenomes were unclassified in the Viral Proteins database, accounting for 45% and 53%, respectively. Of the classified sequences, 42% and 43% were double-stranded DNA (dsDNA) viruses and 13% and 4% were single-stranded DNA (ssDNA) viruses (Table S3.1), in the confined and unconfined aquifers, respectively. Similar findings have been reported in other viral metagenomes, whereby the majority of environmental viral sequences do not match any known sequences in databases (Angly *et al*., 2006; Bench *et al*., 2007; Desnues *et al*., 2008; Rosario *et al*., 2009b). Further, the large number of viral DNA sequences in our dataset was surprising due to the use of a 0.22 µm collection filter, which viruses would be expected to pass through. However, previous metagenomic studies have similarly obtained substantial numbers of virus sequences from samples filtered through 0.22 µm filters (DeLong *et al*., 2006) and their presence in this study likely occurred because filters became clogged by the high levels of fine sediment particles in the samples.

To determine whether groundwater virus communities have intrinsic characteristics, the viral sequences from the confined and unconfined aquifer metagenomes were compared to metagenomes from a variety of other aquatic environments (Table 3.1), using a normalized Goodall's similarity index (Goodall, 1964; 1966) in the MEtaGenome ANalyzer (MEGAN) (Huson *et al*., 2007). Despite geographical proximity, the confined aquifer viral consortia did not resemble those of the unconfined aquifer, and were instead most similar to the viral sequences in the metagenome from a reclaimed water sample, the reusable end-product of wastewater treatment, in Florida (Fig. 3.1) (Rosario *et al*., 2009b; Smith *et al*., 2011; Roudnew *et al*., 2012). This result contradicts the patterns in bacterial taxonomy recently

observed at the same site in South Australia, which showed that the confined aquifer total microbial metagenome, predominantly bacteria, was taxonomically more similar to that of the overlying unconfined aquifer than to any other environment (Smith *et al*., 2011). The lack of similarity between the confined and unconfined aquifer viral communities suggests the viruses were not introduced into the confined aquifer from the overlying unconfined aquifer, indicating the long-term survival of viruses in groundwater.

To identify the taxa contributing to the similarity between the reclaimed water viruses and the confined aquifer viruses, community profiles were generated in MEGAN (Huson *et al*., 2007). The community profile indicated the main taxa contributing to the similarity between the two metagenomes were ssDNA viruses (Fig. 3.2), accounting for 13% and 7% of the viruses in the confined aquifer and reclaimed water, respectively (Fig. 3.2). Within the ssDNA viruses, members of the *Microviridae* dominated, accounting for 55% and 58% in the confined aquifer and reclaimed water source, respectively. In the confined aquifer, members of the *Circoviridae, Geminiviridae* and *Inoviridae* families accounted for 16%, 6% and 4%, respectively, while in the reclaimed water sample, these viral groups accounted for 8%, 5% and 5%, respectively. Unclassified ssDNA viruses comprised 17% and 23% of the ssDNA viruses in the confined aquifer and reclaimed water, respectively. *Nanoviridae* were only found in the confined aquifer sample, accounting for 2% of ssDNA viruses overall (Fig. 3.2 and 3.3). Of the known virus representatives, *Circoviridae, Geminiviridae, Inoviridae, Microviridae* and *Nanoviridae* are all small viruses, with diameters of 7 - 30 nm (Storey *et al*., 1989; Gibbs and Weiller, 1999; Gutierrez *et al*., 2004). Thus, the dominance of these ssDNA viruses is consistent

with the observations that small viruses have the greatest potential for transport through aquifers (Yates, 2000).

Alternatively, in the unconfined aquifer, unclassified ssDNA viruses and members of the *Inoviridae* family accounted for 50% each (Fig. 3.3). *Inoviridae* are filamentous bacteriophage and although they have a small diameter, approximately 7 nm, they have a greater length of approximately 880 nm (Storey *et al*., 1989). As viruses with sizes of 27 – 75 nm are expected to have the greatest potential for spread into deep aquifers (Borchardt *et al*., 2007), the increased abundance of the *Inoviridae* family in the unconfined aquifer suggests the length of these viruses hindered their transport through to deep aquifer systems, when compared to the smaller viruses of the circular *Microviridae*, *Circoviridae*, *Geminiviridae* and *Nanoviridae* families.

Circoviridae, *Geminiviridae* and *Nanoviridae* all contain known plant or vertebrate pathogens (Gibbs and Weiller, 1999; Gutierrez *et al*., 2004). In particular, *Circoviridae* have been characterised from the tissues of birds, mammals, fish, insects, plants, algal cells, and in human and animal faeces (Victoria *et al*., 2009; Delwarta and Li, 2012). Although the origin of circoviruses in human faeces remains unclear (Victoria *et al*., 2009), the broad host range suggests this viral group could be of potential risk to humans. Furthermore, ssDNA viruses are known to have high nucleotide substitution rates, which are thought to contribute to their high pathogenicity and broad host range (Mathews, 2006; Lefeuvre *et al*., 2009). Therefore, the identification of such viruses in this study from a 1500 year-old confined aquifer (Banks *et al*., 2006) suggests the potential exists for long-term survival and spread of small, circular pathogenic viruses in groundwater. Obtaining a broader understanding of potentially pathogenic viral communities within

groundwater is particularly important given the ability of viruses to survive and spread within aquifer ecosystems.

3.3 Acknowledgements

The authors gratefully acknowledge the funding provided by the Australian Research Council. R. J. Smith is the recipient of a Flinders University Research Scholarship (FURS).

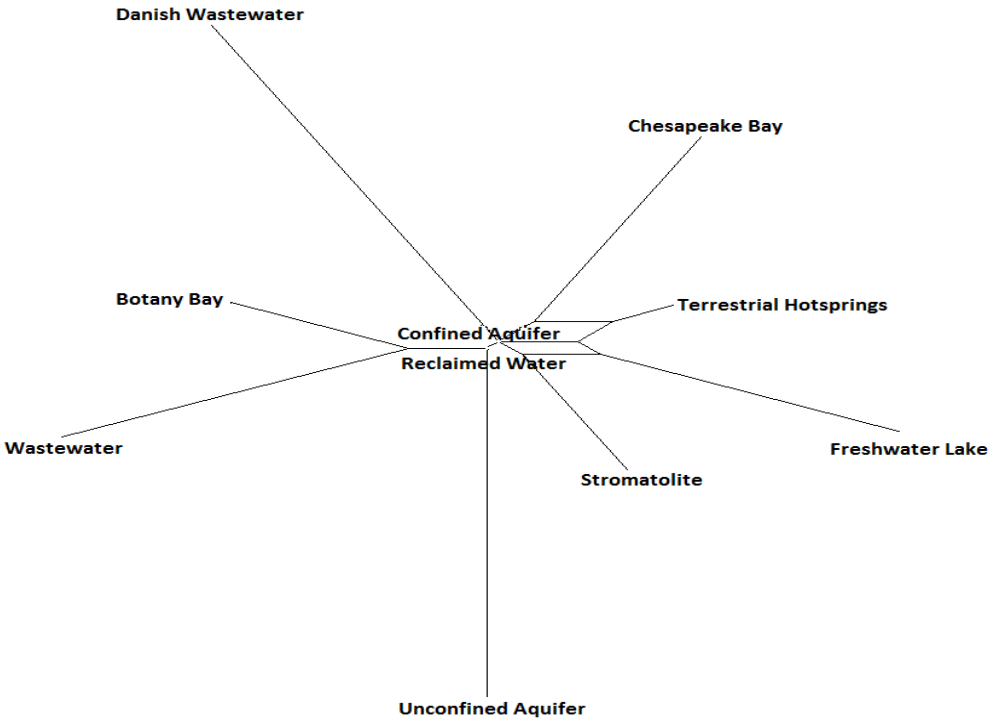

Figure 3.1 Unweighted pairgroup method using arithmetic mean (UPGMA) clustering of viral metagenomes based on normalized Goodall's similarity matrix. Non redundant metagenomic sequences were assembled and identified by using the BLASTX algorithm and $E < 1 \times 10^{-5}$ against the Viral Proteins database using CAMERA (Seshadri *et al*., 2007). Network analysis was then generated from the normalized Goodall's similarity index (Goodall, 1964; 1966) in MEGAN (Huson *et al*., 2007). Goodall's index is designed for determining similarities between multivariate datasets that gives more weight to differences between rare taxa, making it particularly suitable for comparison of microbial metagenomes (Sogin *et al*., 2006; Mitra *et al*., 2010). To visualise relationships between samples, the UPGMA (Sokal and Michener, 1958) clustering was used within MEGAN.

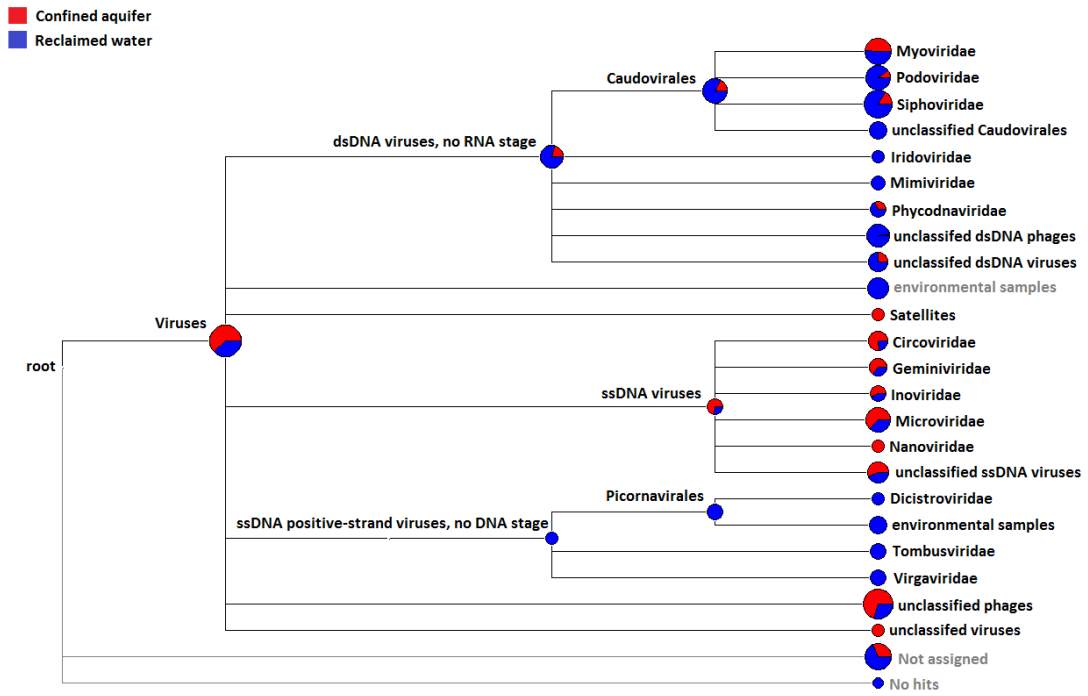

Figure 3.2 Community profile of confined aquifer and reclaimed water metagenomes matching the viral proteins database in CAMERA.

Phyla are expanded to family level where available. Non redundant metagenomic sequences were assembled and identified using the BLASTX algorithm and $E < 1 \times 10^{-5}$ against the Viral Proteins database using CAMERA (Seshadri *et al*., 2007). Normalized abundances were then used to generate a community profile in MEGAN (Huson *et al*., 2007).

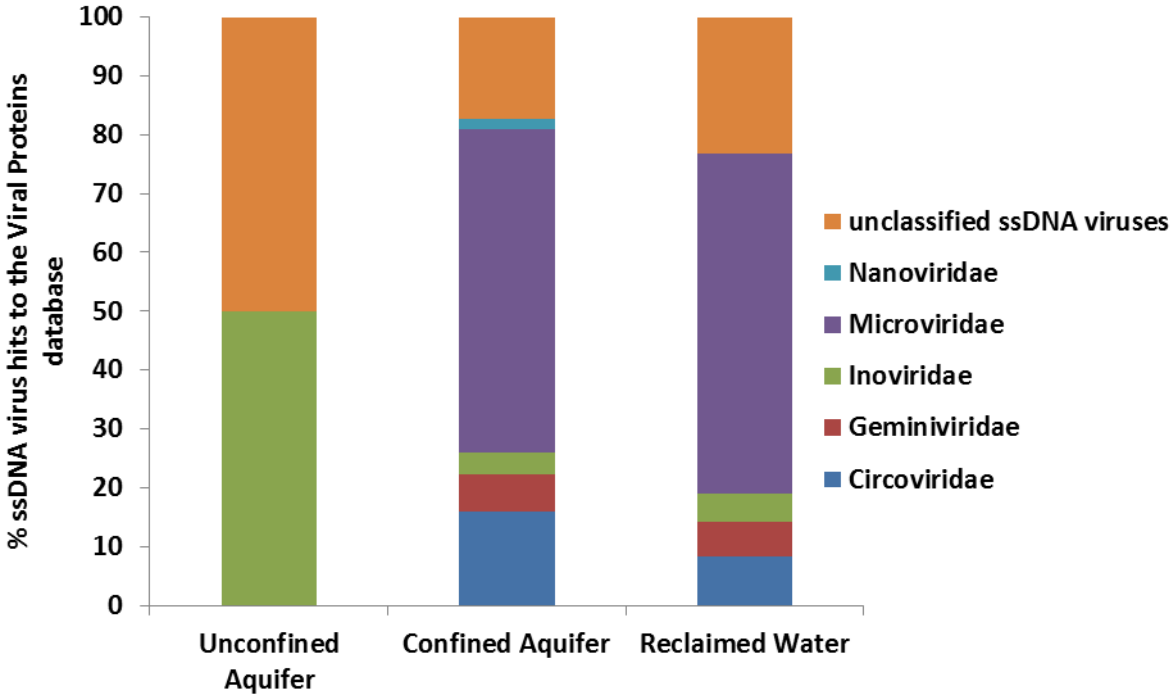

Figure 3.3 ssDNA viruses % relative abundance in the unconfined aquifer, confined aquifer and reclaimed water samples identified by BLASTX to the viral proteins database in CAMERA (Seshadri *et al.*, 2007).

Table S3.1 Relative proportion of matches to the viral proteins database taxonomical hierarchy.

	Confined Aquifer	Unconfined Aquifer
dsDNA viruses, no RNA stage	15.05	15.54
Caudovirales	13.61	14.68
Myoviridae	8.38	7.77
Podoviridae	0.78	0.00
Siphoviridae	3.46	4.89
unclassified Caudovirales	0.00	0.00
Iridoviridae	0.00	0.00
Mimiviridae	0.00	0.00
Phycodnaviridae	0.10	0.00
unclassified dsDNA phages	0.08	0.00
unclassified dsDNA viruses	0.29	0.00
environmental samples	0.00	0.00
Satellites	0.08	0.00
ssDNA viruses	6.77	2.30
Circoviridae	1.04	0.00
Geminiviridae	0.41	0.00
Inoviridae	0.24	0.72
Microviridae	3.55	0.00
Nanoviridae	0.12	0.00
unclassified ssDNA viruses	1.12	0.72
ssRNA viruses	0.00	0.00
ssRNA positive-strand viruses, no DNA stage	0.00	0.00
Picornavirales	0.00	0.00
Dicistroviridae	0.00	0.00
environmental samples+	0.00	0.00
Tombusviridae	0.00	0.00
Virgaviridae	0.00	0.00
unclassified phages	39.89	48.92
unclassified viruses	0.14	0.00
Not assigned	4.89	4.46
No hits	0.00	0.00

Chapter 4

Effect of hydrocarbon impacts on the structure and functionality of marine foreshore microbial communities: A metagenomic analysis

4.0 Abstract

The effect of hydrocarbon contamination on microbial community structure and function was assessed in a historically, hydrocarbon impacted beach sample using metagenomic analysis. Hydrocarbon concentrations of up to 1764 mg kg^{-1} of C_9-C_{36} hydrocarbons were observed at 1.75 m. To assess the effect hydrocarbon impact had on the structure and functionality of foreshore microbial communities, the metagenome from 1.75 m was compared with non-impacted marine metagenomes. A fundamental shift in taxa was seen, with an overrepresentation of *Pseudomonadales*, *Actinomycetales*, *Rhizobiales*, *Alteromonadales*, *Oceanospirillales* and *Burkholderiales* in the hydrocarbon impacted sample. In addition, a relative overrepresentation of metabolic processes including aromatic compound metabolism, nitrogen metabolism and stress response were observed in the hydrocarbon impacted sample. These differences suggest that hydrocarbons in the foreshore environment exerted a selective pressure on microbial consortia, favouring organisms with the ability to catabolise hydrocarbon inputs. Furthermore, power law abundance curves showed the hydrocarbon impacted beach community had mid-range diversity both taxonomically and metabolically, indicative of a functionally redundant and stable community that has adapted to stress. Obtaining a broader understanding of the structure and function of microbial communities inhabiting a historically contaminated site is particularly important given the long term potential persistence and toxicity of hydrocarbon impact.

4.1 Introduction

Hydrocarbons are a ubiquitous class of natural compounds which are found in low concentrations in most soils and sediments (Rosenberg *et al*., 1992; Johnsen and Karlson, 2005). Consequently, hydrocarbon-oxidising microbial communities are present in varying concentrations in the natural environment (Rosenberg, 2006). The presence of hydrocarbon degrading microbial communities have thus, become the source of many studies, due to their potential to clean up contaminants such as hydrocarbons (Chikere *et al*., 2011). Due to their long term persistence and toxicity in the environment (Singleton, 1994), petroleum hydrocarbons have become a common target for bioremediation projects.

Many studies have shown that hydrocarbon contamination can cause a major shift in the structure of microbial communities, with microorganisms capable of surviving and/or utilizing the hydrocarbons as carbon and energy sources becoming dominant (Macnaughton *et al*., 1999; Vinas *et al*., 2005; Wu *et al*., 2008; Kostka *et al*., 2011). These shifts in the microbial community have previously been linked to a reduction in species and allelic diversity within the population, as well as the elimination of some metabolic pathways (Hemme *et al*., 2010). It has been shown that structurally stable microbial communities were less likely to cope with environmental change, due to the inability to retain functionality of the less dominant species, which may contain the genes for bioremediation (Fernandez *et al*., 2000). Thus, flexibility is a major factor contributing to the success of a community to survive, and subsequently degrade contaminants (Marzorati *et al*., 2008).

The rate at which the microbial consortium is able to degrade the contaminant also depends on a variety of environmental factors such as temperature, seasonality and

the availability of nutrients essential for microbial growth (Margesin and Schinner, 2001; Venosa and Zhu, 2003). For example, the degradation of hydrocarbons on sandy beaches is thought to be limited by the availability of inorganic nutrients such as nitrogen and phosphorus (Atlas and Bartha, 1972; Gallego *et al*., 2001; Röling *et al*., 2002), with several studies showing the addition of mineral nutrients significantly enhanced bioremediation (Swannell *et al*., 1995; Venosa *et al*., 1996; Röling *et al*., 2004; Santos *et al*., 2011).

The natural ability of an environmental microbial community to clean up hydrocarbon contamination, without the addition of nutrients, is comparatively less well characterised. Furthermore, information regarding which microorganisms and which functional genes are associated with the catabolism of hydrocarbons is still lacking (Yergeau *et al*., 2012). Advances in high throughput sequencing have allowed for the characterisation of whole environmental microbial communities from the metabolic and taxonomic perspective (Kennedy *et al*., 2010) greatly increasing our potential to understand how indigenous microbial communities respond to hydrocarbon pollution. For example 454 pyrosequencing of hydrocarbon contamination of arctic soils have shown an increase in the abundance of *Alphaproteobacteria* and *Gammaproteobacteria* groups, which are common hydrocarbon degrading groups in contaminated soils (Yergeau *et al*., 2012). Yergeau *et al*., (2012) also found that the abundance of hydrocarbon degrading genes has also been observed to increase due to selective pressure exerted by hydrocarbon pollutants.

Other high throughput sequencing studies have also shown that microbial functional patterns are highly correlated to local environmental factors, with 59% of microbial

community variability explained by oil contamination, geographic location and soil geochemical parameters (Liang *et al*., 2011). Further to this, when oil contaminated beach samples from the Gulf of Mexico were compared to "clean" beach samples, multidimensional scaling plots indicated a uniform response to oil contamination with the oiled samples forming a discrete cluster which was distinct from the clean samples (Kostka *et al*., 2011). Consequently, it is important to build on previous studies by adding detailed metabolic dynamics to general taxonomic presence. Furthermore, the identification of specific degradation and remediation pathways are essential for the understanding of how bacteria remediate hydrocarbons in the natural environment.

The aim of the current study was to utilise next generation metagenomic DNA sequencing to assess the effect of historical hydrocarbon impacts on the taxonomic and metabolic profile of marine ecosystem.

4.2 Materials and Methods

4.2.1 Site selection and sampling

Hydrocarbon contaminated material was sampled from a former oil refinery site in Australia. Approximately 30kg of material was collected from 6 depths (0, 1, 1.25, 1.5, 1.75 and 2 m) at the marine foreshore and subjected to hydrocarbons analysis and microbial community profiling.

4.2.2 Extraction and quantification of hydrocarbon

Hydrocarbons were extracted from samples using an accelerated solvent extractor (ASE200 Accelerated Solvent Extraction System, Dionex Pty Ltd, Lane Cove, NSW, Australia), as previously described by Dandie *et al*., (2010). Freeze-dried ssamples (2-10 g) were ground with diatomaceous earth (Dionex), weighed into extraction

cells and surrogate 100 µl phenanthrene (100 mg ml^{-1}) added prior to sealing. Samples were extracted with hexane:acetone (1:1 v/v) using standard conditions (150 °C, 10.34 MPa, static time 5 min). A steady flow of nitrogen gas was used to concentrate the soil extracts to dryness, and then resuspended in 2 ml of hexane:acetone (1:1 v/v). Prior to analysis, resuspended soil extracts were filtering through 0.45 µm Teflon syringe filters into 2 mL GC vials (Agilent Technologies Australia, Forest Hills, VIC, Australia).

Agilent Technologies 7890A gas chromatograph flame ionisation detector (FID) was used to generate chromatographs of sample extracts. A 15 m x 0.32 mm x 0.1 µm Zebron ZB-5HT (5% phenyl, 95% dimethylpolysiloxane) Inferno column with a 5 m x 0.25 mm inert guard column (Phenomenex Australia, Lane Cove NSW, Australia) was used to separate the samples. Operating conditions were as follows: The oven temperature was programmed at 40 °C for 3 min followed by a linear increase in temperature to 375 °C at 25 °C min^{-1}, and held at 375 °C for 5 min. Detector and injector temperatures were held at 380 °C and 300 °C, respectively. Defined hydrocarbon fractional ranges (C_{6-9}, C_{10-14}, C_{15-28}, C_{29-36}, C_{37-40}) were used to quantify hydrocarbon concentration using Window defining standards (Accustandard Inc., New Haven, CT USA). Hydrocarbon concentrations were quantified according to Dandie *et al*., (2010) and reported per g freeze-dried sample. Surrogate recovery during hydrocarbon quantification ranged from 94-103%, while results of replicate analysis of the same sample showed a standard deviation of less than 8%.

4.2.3 Nutrient analysis, microbial community DNA extraction and sequencing for metagenomic analysis

Based on hydrocarbon profiling results, samples from a depth of 1.75 m were subjected to metagenomic analysis. Triplicate samples (30 g) were collected and stored on ice following collection for physiochemical characterisation. Nutrient analysis for total nitrogen and total phosphorus were conducted using a segmented flow analyser and colorimetric techniques (APHA, 2005). All analysis was conducted at the Australian Water Quality Centre (Adelaide).

Following collection, microbial community DNA was extracted using the PowerMax® Soil DNA Isolation Kit (MoBio laboratories, Inc., Carlsbad, CA, USA). DNA quality and concentration was then determined by 1.5% TBE agarose gel electrophoresis (Bioline) and a Qubit fluorometer (Quant-iT^{TM} dsDNA HS Assay Kit; Invitrogen Inc.). Approximately 500 ng of high molecular weight DNA was then sequenced on the GS-FLX pyrosequencing platform using Titanium series reagents (Roche) at the Ramaciotti Center for Gene Function Analysis, Sydney, Australia.

4.2.4 Data analysis

Annotation of the unassembled DNA sequences was performed with the MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) pipeline version 3.0 (Meyer *et al*., 2008). BLASTX was performed with an E-value cut-off of $E<1e^{-5}$ and a minimum alignment length of 50 bp as described by Dinsdale *et al*. (2008b). Metabolic profiles were produced using the normalized abundance of sequence matches to a given subsystem, while the normalized abundance of sequence matches to the SEED database (http://metagenomics.theseed.org/) (Overbeek *et al*., 2005) were used to generate taxonomic profiles.

The metagenome from the hydrocarbon impacted foreshore were compared to non-impacted foreshore sediment from Jeffries *et al*. (2011a) (Table S4.1). These metagenomes were sampled from two different locations nearby the study site, providing a baseline for which the hydrocarbon impacted foreshore could be compared. Furthermore, the use of two sites allowed for any bias that may have been apparent due to difference in location to be reduced. The Statistical Analysis of Metagenomic Profiles (STAMP) software package was used to determine the statistically significant differences between the hydrocarbon impacted and non-impacted sites (Parks and Beiko, 2010). Firstly, a frequency table of the number of hits to each individual taxa or subsystem for each metagenome was generated using an E-value cut-off of $E<1e^{-5}$ to identify hits. To remove bias in difference in read lengths and sequencing effort, the frequency table was normalised by dividing by the total number of hits. *P*-values were calculated in STAMP using the two sided Fisher's Exact test (Fisher, 1958), while confidence intervals were calculated using the Newcombe-Wilson method (Newcombe, 1998). False discovery rate was corrected for using the Benjamini-Hochberg FDR method (Benjamini and Hochberg, 1995). To avoid bias based on location, only those that were found to be overrepresented when compared to both controls were included for discussion. The main subsystems contributing to the differences between community structure were identified using similarity percentage (SIMPER) analysis (Clarke, 1993).

To determine the overall influence hydrocarbon impact had on the microbial communities both structurally and functionally, rank abundance plots were generated and compared to the metagenomes from 9 other marine environments (Table S4.1). Frequency tables were generated in MG-RAST as above. Taxa/metabolism rank was plotted on the x-axis and the relative abundance was plotted on the y-axis, where had

both been log_{10} transformed. The noise/rare biosphere was left out as per Mitchell (2004). The data that produced the best fit had a power law trend line assigned.

4.3 Results

4.3.1 Nutrient and hydrocarbon analysis

Samples were collected during test pit activities at the marine foreshore with bulk samples collected at ground surface and from depths of 0, 1.0, 1.25, 1.5, 1.75 and 2.0 m. Hydrocarbon concentrations were below the level of quantification in surface samples and samples collected at 0, 1.0, 1.25 and 1.5 m. However, C_6-C_9, C_{10}-C_{14} and C_{15}-C_{28} hydrocarbon fractional ranges were detected at 1.75 and 2.0 m. In samples collected from 1.75 and 2.0 m, low level C_6-C_{36} hydrocarbon concentrations (Sheppard *et al.*, 2011) of 1764 and 1420 mg kg^{-1} respectively were observed, with the concentrations predominantly composed of the C_{15}-C_{28} hydrocarbons (Table 4.1). Total soil nitrogen and phosphorus concentrations were low throughout the depth profile with maximum concentrations of 55 and 40 mg kg^{-1} at 1.75 m, respectively (Table 4.1).

4.3.2 Taxonomic and metabolic profiling of beach metagenomes

A total of 229,089 sequences with an average read length of 424 bases were obtained from the hydrocarbon impacted foreshore sample. The hydrocarbon impacted foreshore metagenomic library was 92.5% bacteria, by SEED database matches. Proteobacteria represented 69.5% bacterial matches, and within this, Gammaproteobacteria contributed to 31.8% of matches in the hydrocarbon impacted foreshore sample. A total of 6.3% reads could not be assigned to any known sequence in the database (Table S4.2). The remainder of the sequence matches were Archaea (0.9%), Eukaryota (0.4%) and Viruses (0.02%).

Table 4.1 Properties of samples used in this study

	Hydrocarbon (mg kg^{-1})					
Constituent	**0 m**	**1.0 m**	**1.25 m**	**1.5 m**	**1.75 m[a]**	**2.0 m**
BTEX	< LOR[b]	<LOR	< LOR	< LOR	< LOR	< LOR
C_6-C_9	< LOR[c]	<LOR	< LOR	< LOR	34	20
C_{10}-C_{14}	< LOR[d]	<LOR	< LOR	< LOR	500	360
C_{15}-C_{28}	< LOR[e]	<LOR	< LOR	< LOR	1230	1040
C_{29}-C_{36}	< LOR[f]	<LOR	< LOR	< LOR	< LOR	< LOR

[a] Total Nitrogen and Total Phosphorus at a depth of 1.75m were 55.0 ± 0.0 and 40.3 ± 6.0, respectively.
[b]Level of reporting for toluene, ethylbenzene and xylene was 0.5 mg kg-1 and 0.2 mg kg^{-1} for benzene.
[c]Level of reporting for C_6-C_9 hydrocarbons was 10 mg kg^{-1}.
[d]Level of reporting for C_{10}-C_{14} hydrocarbons was 50 mg kg^{-1}.
[e]Level of reporting for C_{15}-C_{28} hydrocarbons was 100 mg kg^{-1}.
[f]Level of reporting for C_{29}-C_{36} hydrocarbons was 100 mg kg^{-1}.

Differences were observed between the hydrocarbon impacted foreshore sample when compared to two non-impacted foreshore samples using STAMP. An overrepresentation of Proteobacteria and Actinobacteria were seen in the hydrocarbon impacted foreshore sample. Conversely, there was an overrepresentation of Cyanobacteria, Bacteroidetes, Planctomycetes, Acidobacteria and Firmicutes in both non-impacted samples (*q*-value $<1e^{-15}$) (Fig. 4.1). At the order level of taxonomic resolution, *Pseudomonadales*, *Actinomycetales*, *Rhizobiales*, *Alteromonadales*, *Oceanospirillales* and *Burkholderiales* were overrepresented in the hydrocarbon impacted sample while, *Planctomycetales*, *Flavobactriales*, *Desulfobacterales*, *Nostocales*, *Rhodobacterales*, *Bacteroidales*, and *Cytophagales* were overrepresented in the non-impacted samples (*q*-value $<1e^{-15}$) (Fig. 4.2).

The core metabolic function in the hydrocarbon impacted foreshore sample was carbohydrate metabolism, while a high level of biotin biosynthesis, metabolism of fatty acids and aromatic compound catabolism was also observed. Within this, the highest pathway contributing to aromatic compound metabolism was n-Phenylalkanoic acid degradation and anaerobic benzoate degradation (Table S4.3). Comparisons of metabolic profiles for impacted and non-impacted samples using STAMP revealed an overrepresentation of genes corresponding to nitrogen metabolism, stress response and aromatic compound metabolism in the impacted foreshore sample. Alternatively, carbohydrate metabolism was overrepresented in the non-impacted samples (*q*-value $<1e^{-5}$) (Fig. 4.3). Further to this, SIMPER analysis revealed that the metabolism of aromatic compounds genes (higher in the impacted sample) and motility and chemotaxis genes (higher in the non-impacted samples) accounted for the majority of the dissimilarity between the impacted and non-impacted samples (Table S4.4 and S4.5).

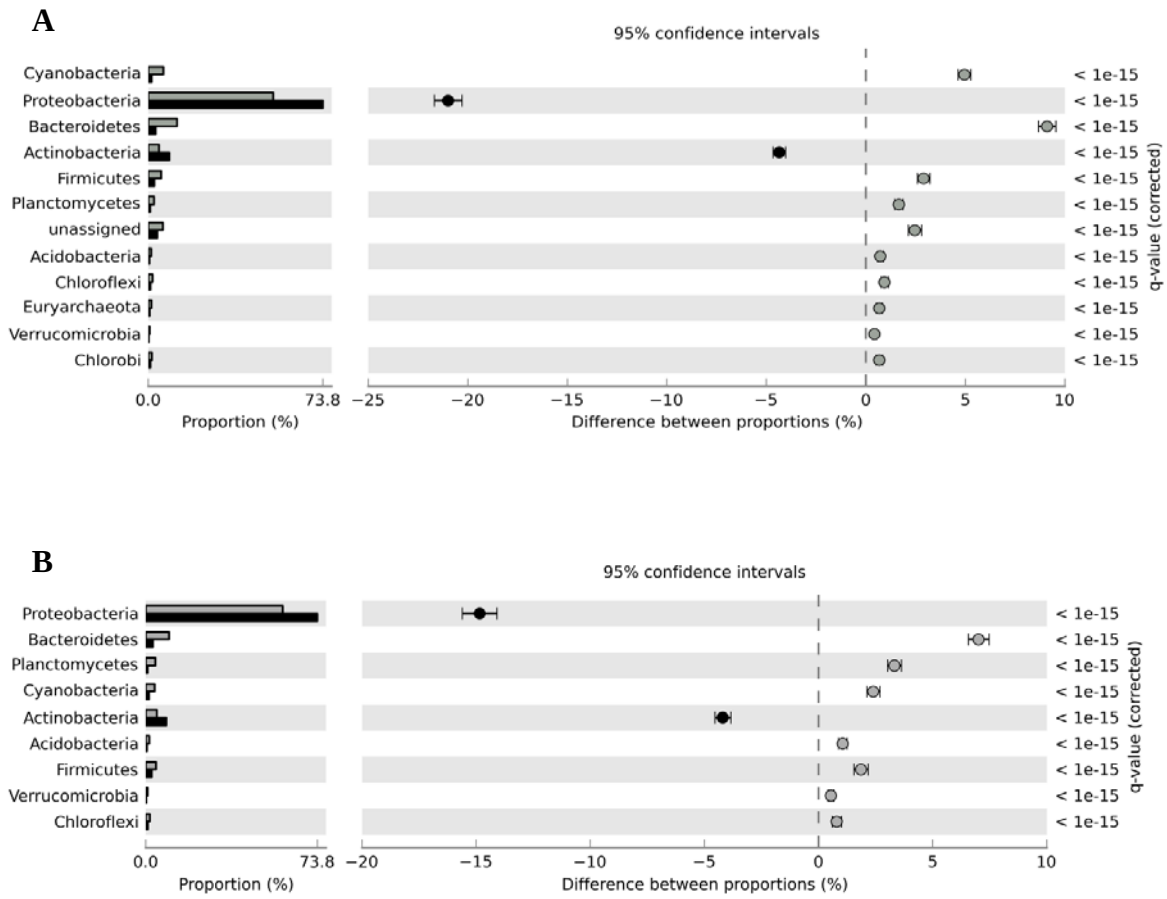

Figure 4.1 Comparison of foreshore taxonomic profiles at phylum level: (A) STAMP analysis of taxonomy enriched or depleted between the hydrocarbon-impacted foreshore sample and non-impacted marine sample 1. Groups overrepresented in non-impacted sample 1 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Benjamini-Hochberg FDR. A *q*-value cut-off of $<1e^{-15}$ was then implemented. (B) STAMP analysis of taxonomy enriched or depleted between the hydrocarbon-impacted foreshore samples and non-impacted marine sample 2. Groups overrepresented in non-impacted sample 2 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions.

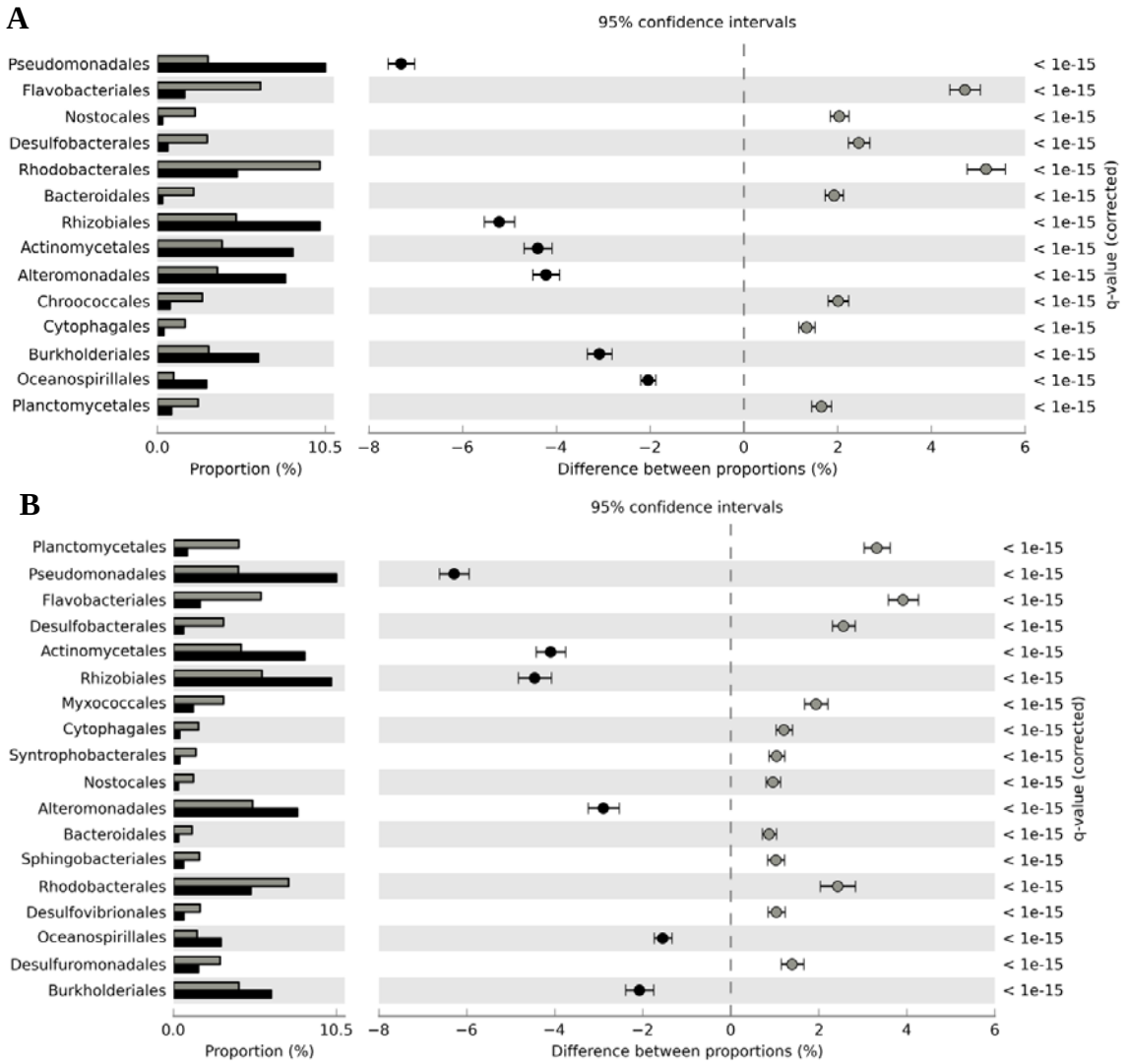

Figure 4.2 Comparison of foreshore taxonomic profiles at order level taxonomy: (A) STAMP analysis of taxonomy enriched or depleted between the hydrocarbon-impacted foreshore sample and non-impacted marine sample 1. Groups overrepresented in non-impacted sample 1 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Benjamini-Hochberg FDR. A *q*-value cut-off of $<1e^{-15}$ was then implemented. (B) STAMP analysis of taxonomy enriched or depleted between the hydrocarbon-impacted foreshore sample and non-impacted sample 2. Groups overrepresented in non-impacted sample 2 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions.

To determine the overall effect hydrocarbon impact had on the diversity of the microbial community, both in terms of structure and function, we compared the hydrocarbon impacted foreshore sample with 9 publicly available metagenomes on MG-RAST from a variety of marine environments (Table S4.1). The highest metabolic (subsystem) and taxonomic (organism) resolution available was used to create rank abundance curves. Analysis of the slope of the power law fits to rank abundance plots revealed a community with mid-range distribution (λ= -0.411 and -540 for taxonomy and metabolism, respectively), which was similar to those from other oligotrophic marine environments (Table 4.2).

4.4 Discussion

Effective bioremediation in marine environments is known to be limited by factors such as nutrient availability, temperature and oxygen concentration (Röling *et al*., 2002; Kostka *et al*., 2011). Many studies have focused on the taxonomic shifts hydrocarbons exert on coastal marine microbial communities (Chikere *et al*., 2011; Liang *et al*., 2011; Yergeau *et al*., 2012), however, the pathways by which bioremediation of hydrocarbons is achieved in these environments, as well as the long term persistence of such pathways, is still relatively unknown. To determine the long term effect hydrocarbon impacts have on microbes in marine foreshore environments, the microbial ecology of a historically impacted site was assessed to determine the influence on microbial taxonomy and metabolism.

Vertical profiling of hydrocarbon impacted foreshore samples over 0 – 2.0 m showed elevated hydrocarbon concentrations of up to 1764 mg kg^{-1} of C_9-C_{36} hydrocarbons at 1.75 m (Table 4.1). This is consistent with other reports that have shown hydrocarbon concentrations may be elevated in the sub-surface marine environments

(Ke *et al*., 2005) as a result of vertical transport by tidal action (Röling *et al*., 2004). This may result in recalcitrant hydrocarbon fractions of crude oil persisting in sub-surface environments (Short *et al*., 2007).

To determine how hydrocarbon impacts influence indigenous microbial communities within a marine environment, we compared our metagenome to two other metagenomes obtained from non-hydrocarbon impacted marine foreshore sediment (Jeffries *et al*., 2011a). Differences were observed between the hydrocarbon impacted sample compared to the non-impacted samples, with a shift in dominant taxa between the impacted and non-impacted samples, suggesting markedly different community compositions. In the hydrocarbon impacted foreshore sample, there was an overrepresentation of *Pseudomonadales*, *Actinomycetales*, *Rhizobiales*, *Alteromonadales*, *Oceanospirillales* and *Burkholderiales* (Fig. 4.2). These findings are similar to those reported by Marcial Gomes *et al*., (2008) who used 16S rRNA sequencing to show that there was an enrichment in ribotypes related to *Alteromonadales*, *Burkholderiales*, *Pseudomonadales*, *Rhodobacterales* and *Rhodocyclales* in urban mangrove forest sediments polluted with hydrocarbons. Thus, the overrepresentation of such groups within the hydrocarbon impacted foreshore metagenome, suggests that the innate potential exists within the microbial consortium inhabiting this environment, for the degradation of hydrocarbons.

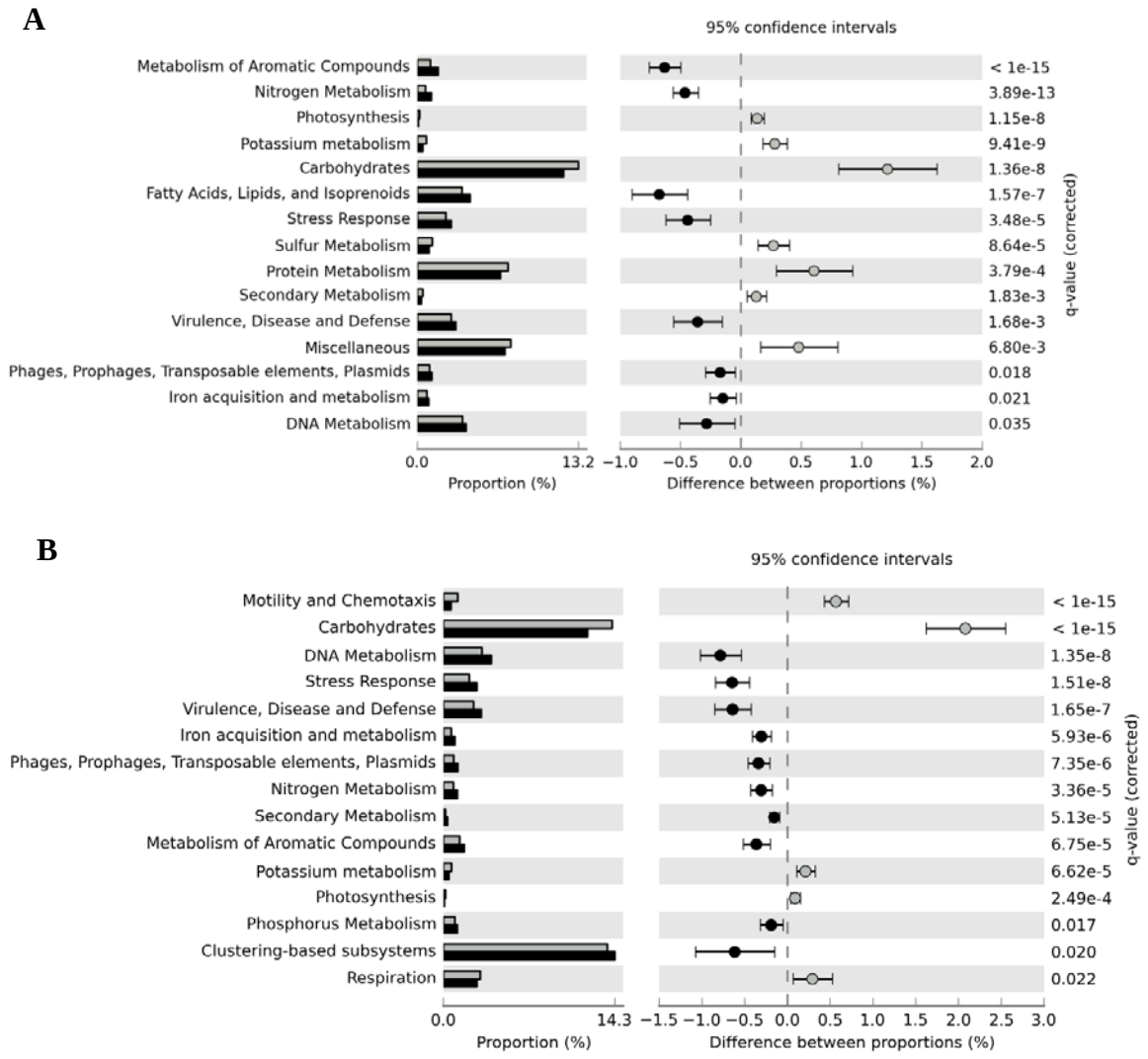

Figure 4.3 Comparison of foreshore metabolic profiles, hierarchy level 1: (A) STAMP analysis of metabolisms enriched or depleted between the hydrocarbon-impacted foreshore sample and non-impacted marine sample 1. Groups overrepresented in non-impacted sample 1 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions. Corrected *P*-values (*q*-values) were calculated using Benjamini-Hochberg FDR. (B) STAMP analysis of metabolism enriched or depleted between the hydrocarbon-impacted foreshore sample and non-impacted sample 2. Groups overrepresented in non-impacted sample 2 (grey) correspond to positive differences between proportions and groups overrepresented in the hydrocarbon-impacted foreshore sample (black) correspond to negative differences between proportions.

The overrepresentation of *Oceanospirillales* in the hydrocarbon impacted foreshore sample is notable due to this species' ability to dominate in hydrocarbon impacted marine environments (Hazen *et al*., 2010; Atlas and Hazen, 2011). This success has previously been linked to their ability to degrade branched chain alkanes, like those found in this study (Table 4.1), thus outcompeting other associated microorganisms (Hara *et al*., 2003). *Oceanospirillales* spp. are known to produce biosurfactants which aids in the emulsification of alkanes, by increasing their bioavailability and thus, increasing the rate of degradation (Schneiker *et al*., 2006). In addition, *Oceanospirillales* spp. have also been shown to proliferate in an oligotrophic marine environment due to their innate ability to effectively scavenge key elements such as nitrogen and phosphorus (Martins dos Santos *et al*., 2010). This enables them to quickly and effectively adapt to sudden increases in carbon and the corresponding decreases of other nutrients such as nitrogen and phosphorus following hydrocarbon utilisation (Schneiker *et al*., 2006). Furthermore, as *Oceanospirillales* are generally associated with marine environments, their overrepresentation in the hydrocarbon contaminated beach sample suggests the microbial potential to degrade hydrocarbons is being enhanced by selective pressure favouring these species, as well as coastal/seawater interactions, which are consequently introducing microbes possessing the capacity to catabolise hydrocarbons.

The rate at which microbial communities are able to biodegrade hydrocarbons in the environment is dependent on nitrogen, phosphorus and hydrocarbon bioavailability (Nikolopoulou and Kalogerakis, 2008), in addition to the presence and expression of genes responsible for their catabolism. In marine foreshore environments, nutrients concentrations are generally thought to be too low for successful bioremediation (Röling *et al*., 2002). In this study, nutrient analysis of hydrocarbon impacted

samples also showed low nitrogen and phosphorus concentrations (55 mg kg^{-1} and 40 mg kg^{-1} respectively) (Table 4.1). Further evidence of this is the detection of microbes such as the *Oceanospirillales* spp., which are known for their ability to successfully scavenge nutrients in low concentrations. The overrepresentation of nitrogen metabolism genes in the hydrocarbon impacted foreshore sample suggests scavenging mechanisms may be in place where nitrogen concentrations are paramount for hydrocarbon catabolism compared to low carbon, non-impacted environments (Fig. 4.3).

Our data also indicated that aromatic hydrocarbon metabolism genes were overrepresented in the hydrocarbon impacted foreshore sample (Fig. 4.3), with n-Phenylalkanoic acid degradation genes being the most abundant (Table S4.3). Previous studies have demonstrated the ability for *Pseudomonas* spp. to metabolise phenylalkanoic acids, a component of polyhydroxyalkanoate (PHA) found in crude oil (Sabirova, 2010). These compounds are used as an intracellular carbon storage material in response to excess carbon and nutrient deficiencies (Madison and Huisman, 1999). Hydrocarbon degradation genes are widely distributed in marine environments (Head *et al*., 2006). In pristine sites, microbes capable of degrading hydrocarbons are thought to utilize natural sources such as those produced by algae, plants and other organisms (Atlas, 1995; Yergeau *et al*., 2012). Following hydrocarbon contamination, there is an increase in the proportion of microbial populations with plasmids containing genes for hydrocarbon degradation (Leahy and Colwell, 1990; Atlas, 1995). The abundance of n-Phenylalkanoic acid degradation genes in the oligotrophic hydrocarbon impacted foreshore sample is, therefore consistent with the ability to catabolise petroleum hydrocarbons under low nutrient conditions.

Anaerobic benzoate degradation genes were also present in the hydrocarbon impacted foreshore sample (Table S4.3). Although the concentration of BTEX were below the level of quantification at the time of this study, aromatic hydrocarbons may have been present during the initial impact and were probably degraded over time nearer ground surface due to reduced oxygen tension. Benzene degradation is known to be impaired by anaerobic conditions (Holmes *et al*., 2011) although reports by van der Zaan *et al*., (2012) have shown that degradation of aromatic compounds can occur, albeit a slower rate compared to aerobic conditions. Previous exposure of samples at these depths to aromatic hydrocarbons could, therefore, have played a role in the abundance of these genes. The presence of anaerobic benzoate degradation genes along with the n-Phenylalkanoic acid degradation genes indicates that the adaptation of microbial communities to hydrocarbon impacts can remain for long periods of time, whereby years later, the community is still typical of communities responding to a recent contaminated event.

Table 4.2 Comparison of microbial community evenness and functional stability in marine environments. Power distribution with exponents (λ)

	Taxonomy		Metabolism	
Metagenome	**λ**	**R^2**	**λ**	**R^2**
Coastal Galapagos Island	-0.288	0.968	-0.743	0.958
East Australian Current 1	-0.296	0.979	-0.738	0.958
Botany Bay	-0.300	0.987	-0.843	0.936
East Australian Current 2	-0.306	0.932	-0.642	0.941
Lagoon Reef - Indian Ocean	-0.319	0.972	-0.838	0.953
Marine Sediment 1 (non-impacted)	-0.385	0.939	-0.500	0.980
Marine Sediment 2 (non-impacted)	-0.386	0.978	-0.497	0.961
HOT 10m	-0.409	0.952	-0.576	0.952
Hydrocarbon impacted beach	**-0.411**	**0.991**	**-0.540**	**0.986**
HOT 200m	-0.420	0.977	-0.533	0.935

To determine how the historical contamination event influenced the overall structural and functional dynamics of the microbial community, we compared the metagenome from the hydrocarbon impacted foreshore with metagenomes from 9 other marine habitats (Table S4.1). Taxonomically and metabolically, the hydrocarbon impacted foreshore exhibited mid-range diversity (λ= -0.411 and -540, respectively) indicative of a bacterial community, which is likely to have adapted to stress (Table 4.2). Such communities possess sufficient functional redundancy allowing for community evenness and functional organization to remain stable, and largely unaffected by environmental stress (Marzorati *et al*., 2008). The initial hydrocarbon impact at the study site occurred at ground surface with hydrocarbons subsequently transported through the foreshore profile resulting in the accumulation at the sand-bedrock interface. In addition, these beach samples were subjected to constant input of nutrients and water from tidal and wave action, as well as low level contact with contaminants in sea water. This influx is likely to keep the relevant degradation genes selected for and induced, thus resulting in a functionally redundant community.

In conclusion, our data revealed the taxa and functional genes responsible for the catabolism of hydrocarbon in a historically impacted foreshore. The overrepresentation of *Pseudomonadales*, *Burkholderiales* and *Oceanospirillales* as well as nitrogen metabolism genes and aromatic hydrocarbon metabolism genes such as n-Phenylalkanoic acid degradation and anaerobic benzoate degradation in the hydrocarbon impacted foreshore metagenome are all consistent with the bioremediation of hydrocarbons. We suggest this pattern is driven by the coastal/seawater interactions which have created a nutrient flux as well as hydrocarbon degrading marine bacteria. Our data also revealed a functionally

redundant community suggesting that the indigenous microbial communities have adapted and flourished following the initial impact. With the use of next generation sequencing protocols, this study provides important insights into a microbial community's innate ability to degrade hydrocarbons in a naturally low nutrient environment.

4.6 Acknowledgements

R. J. Smith is the recipient of a Flinders University Research Scholarship (FURS).The authors gratefully acknowledge the funding provided by ARC linkage Grant LP0776478 for the metagenomic analysis. Hydrocarbon impacted foreshore sampling and chemical analysis was funded by the Cooperative Research Centre for Contamination Assessment and Remediation of the Environment (CRC CARE), grant number 6-5-01. The authors would like to acknowledge the support of the School of Biological Sciences, Flinders University, the Plant Functional Biology and Climate Change Cluster, University of Technology Sydney and the Centre for Environmental Risk Assessment and Remediation, University of South Australia.

Table S4.1 Summary of metagenomes used in this study

MG-RAST ID	Description/Reference
4453082.3	**Hydrocarbon impacted foreshore**
4446341.3	Non-impacted foreshore sediment 1 (Jeffries *et al.*, 2011a)
4446342.3	Non-impacted foreshore sediment 2 (Jeffries *et al.*, 2011a)
4443688.3	Botany Bay 1 (Burke *et al*., 2011)
4446457.3	East Australian Current 1 (Seymour *et al*., 2012)
4446409.3	East Australian Current 2 (Seymour *et al*., 2012)
4441595.3	Coastal Galapagos Island (Rusch *et al*., 2007)
4441139.3	Lagoon Reef - Indian Ocean (Rusch *et al*., 2007)
4441051.3	HOT station 10m (DeLong *et al*., 2006)
4441041.3	HOT station 200m (DeLong *et al*., 2006)

Table S4.2 Relative proportion of matches to the SEED database taxonomic hierarchy

Domain	MG-RAST Level 2 (Phyla)	MG-RAST Level 3 (Class)	Contaminated Beach
Bacteria	Proteobacteria	Gammaproteobacteria	31.758
Bacteria	Proteobacteria	Alphaproteobacteria	22.169
Bacteria	Actinobacteria	Actinobacteria (class)	10.285
Bacteria	Proteobacteria	Betaproteobacteria	9.811
Bacteria	Proteobacteria	Deltaproteobacteria	5.028
unassigned	unassigned	unassigned	4.144
Bacteria	Bacteroidetes	Flavobacteria	2.202
Bacteria	Firmicutes	Clostridia	1.744
Bacteria	Cyanobacteria	unclassified (derived from Cyanobacteria)	1.459
Bacteria	Firmicutes	Bacilli	1.404
Bacteria	Chlorobi	Chlorobia	1.053
Bacteria	Planctomycetes	Planctomycetacia	0.995
Bacteria	Deinococcus-Thermus	Deinococci	0.986
Bacteria	Bacteroidetes	Sphingobacteria	0.777
Bacteria	Chloroflexi	Chloroflexi (class)	0.759
Bacteria	Bacteroidetes	Cytophagia	0.527
Bacteria	Proteobacteria	Epsilonproteobacteria	0.420
Bacteria	Bacteroidetes	Bacteroidia	0.410
Bacteria	Acidobacteria	Solibacteres	0.391
Archaea	Euryarchaeota	Methanomicrobia	0.374
Bacteria	Proteobacteria	unclassified (derived from Proteobacteria)	0.297
Bacteria	Chloroflexi	Thermomicrobia (class)	0.279
Bacteria	Verrucomicrobia	Opitutae	0.199
Bacteria	Acidobacteria	unclassified (derived from Acidobacteria)	0.191
Archaea	Euryarchaeota	Halobacteria	0.186
Bacteria	Thermotogae	Thermotogae (class)	0.174
Bacteria	Cyanobacteria	Gloeobacteria	0.140
Bacteria	Spirochaetes	Spirochaetes (class)	0.132
Bacteria	unclassified (derived from Bacteria)	unclassified (derived from Bacteria)	0.120
Bacteria	Synergistetes	Synergistia	0.118
Bacteria	Aquificae	Aquificae (class)	0.107
Archaea	Crenarchaeota	Thermoprotei	0.104
Eukaryota	Arthropoda	Insecta	0.095
Bacteria	Chlamydiae	Chlamydiae (class)	0.088
Bacteria	Dictyoglomi	Dictyoglomia	0.083

Bacteria	Chloroflexi	Dehalococcoidetes	0.079
Bacteria	Deferribacteres	Deferribacteres (class)	0.074
Eukaryota	Streptophyta	unclassified (derived from Streptophyta)	0.074
Bacteria	Verrucomicrobia	unclassified (derived from Verrucomicrobia)	0.071
Bacteria	Fusobacteria	Fusobacteria (class)	0.065
Archaea	Euryarchaeota	Thermococci	0.064
Eukaryota	Chordata	Mammalia	0.057
Eukaryota	Ascomycota	Sordariomycetes	0.056
Bacteria	Verrucomicrobia	Verrucomicrobiae	0.042
Eukaryota	Chordata	Actinopterygii	0.040
Archaea	Euryarchaeota	Methanococci	0.034
Archaea	Euryarchaeota	Archaeoglobi	0.031
Bacteria	Tenericutes	Mollicutes	0.030
Viruses	unclassified (derived from Viruses)	unclassified (derived from Viruses)	0.026
Bacteria	Elusimicrobia	Elusimicrobia (class)	0.023

Top 50 hits were generated by BLASTing sequences to the MG-RAST subsystem database with a minimum alignment length of 50 bp and an E-value cut-off of $1e^{-5}$.

Relative representation in the metagenome was calculated by dividing the number of hits to each category by the total number of hits to all categories.

Table S4.3 Relative proportion of matches to the subsystem database metabolic hierarchy

Subsystem Hierarchy 1	Subsystem Hierarchy 2	Subsystem Hierarchy 3	% hits
Carbohydrates	One-carbon Metabolism	Serine-glyoxylate cycle	0.3543
Cofactors, Vitamins, Prosthetic Groups, Pigments	Biotin	Biotin biosynthesis	0.2975
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty acid degradation regulons	0.2975
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty acid metabolism cluster	0.2975
Metabolism of Aromatic Compounds	Peripheral pathways for catabolism of aromatic compounds	n-Phenylalkanoic acid degradation	0.2975
Iron acquisition and metabolism	Iron acquisition in Vibrio	-	0.2207
Membrane Transport	Ton and Tol transport systems	-	0.2207
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	0.2095
Clustering-based subsystems	CBSS-235.1.peg.567	-	0.2087
Clustering-based subsystems	Biosynthesis of galactoglycans and related lipopolysacharides	CBSS-258594.1.peg.3339	0.2023
Miscellaneous	Plant-Prokaryote DOE project	COG0451	0.2023
Amino Acids and Derivatives	Branched-chain amino acids	Isoleucine degradation	0.1792
Amino Acids and Derivatives	Branched-chain amino acids	Valine degradation	0.1792
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	0.1792
Carbohydrates	Fermentation	Butanol Biosynthesis	0.1792
Clustering-based subsystems	Butyrate metabolism cluster	-	0.1792
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty acid degradation regulons	0.1792
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty acid metabolism cluster	0.1792
Fatty Acids, Lipids, and	Polyhydroxybutyrate metabolism	-	0.1792

Isoprenoids			
Metabolism of Aromatic Compounds	Peripheral pathways for catabolism of aromatic compounds	n-Phenylalkanoic acid degradation	0.1792
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Multidrug Resistance Efflux Pumps	0.1720
Cell Wall and Capsule	Cell wall of Mycobacteria	mycolic acid synthesis	0.1688
Clustering-based subsystems	Fatty acid metabolic cluster	CBSS-246196.1.peg.364	0.1688
Clustering-based subsystems	Fatty acid metabolic cluster	COG1399	0.1688
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty Acid Biosynthesis FASII	0.1688
Clustering-based subsystems	CBSS-196620.1.peg.2477	-	0.1600
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	BlaR1 Family Regulatory Sensor-transducer Disambiguation	0.1600
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Copper homeostasis	0.1600
Amino Acids and Derivatives	Branched-chain amino acids	Isoleucine degradation	0.1568
Amino Acids and Derivatives	Branched-chain amino acids	Valine degradation	0.1568
Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Lysine fermentation	0.1568
Carbohydrates	Fermentation	Acetone Butanol Ethanol Synthesis	0.1568
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	0.1568
Carbohydrates	Fermentation	Butanol Biosynthesis	0.1568
Carbohydrates	Organic acids	Isobutyryl-CoA to Propionyl-CoA Module	0.1568
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	5-FCL-like protein	0.1568
Fatty Acids, Lipids, and Isoprenoids	Fatty acids	Fatty acid degradation regulons	0.1568
Respiration	Electron accepting reactions	Anaerobic respiratory reductases	0.1568
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	0.1480
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	YgfZ	0.1448

Respiration	Soluble cytochromes and functionally related electron carriers	-	0.1448
Sulfur Metabolism	Inorganic sulfur assimilation	Inorganic Sulfur Assimilation	0.1448
Virulence, Disease and Defense	Resistance to antibiotics and toxic compounds	Cobalt-zinc-cadmium resistance	0.1408
Clustering-based subsystems	CBSS-350688.3.peg.1509	-	0.1360
DNA Metabolism	DNA replication	DNA-replication	0.1360
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage regulation of gene expression	0.1352
Stress Response	Oxidative stress	Regulation of Oxidative Stress Response	0.1352
Carbohydrates	Central carbohydrate metabolism	Methylglyoxal Metabolism	0.1176
Carbohydrates	Central carbohydrate metabolism	Pyruvate metabolism II: acetyl-CoA, acetogenesis from pyruvate	0.1176
Fatty Acids, Lipids, and Isoprenoids	Phospholipids	Glycerolipid and Glycerophospholipid Metabolism in Bacteria	0.1176
Miscellaneous	Plant-Prokaryote DOE project	DOE COG2016	0.1152
Protein Metabolism	Selenoproteins	Glycine reductase, sarcosine reductase and betaine reductase	0.1152
Amino Acids and Derivatives	Glutamine, glutamate, aspartate, asparagine; ammonia assimilation	Aspartate aminotransferase	0.1080
Amino Acids and Derivatives	Glutamine, glutamate, aspartate, asparagine; ammonia assimilation	Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis	0.1080
Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Threonine and Homoserine Biosynthesis	0.1080
Miscellaneous	Plant-Prokaryote DOE project	PROSC	0.1080
DNA Metabolism	DNA repair	DNA repair, UvrABC system	0.1072
Amino Acids and Derivatives	Alanine, serine, and glycine	Alanine biosynthesis	0.1072
Clustering-based subsystems	Cell Division	CBSS-393130.3.peg.794	0.1072
Clustering-based subsystems	Lysine, threonine, methionine, and cysteine	CBSS-84588.1.peg.1247	0.1072
Cofactors, Vitamins, Prosthetic Groups, Pigments	Folate and pterines	YgfZ	0.1072

Miscellaneous	Plant-Prokaryote DOE project	At5g37530	0.1072
Miscellaneous	Plant-Prokaryote DOE project	COG2363	0.1072
Miscellaneous	Plant-Prokaryote DOE project	Iron-sulfur cluster assembly	0.1072
RNA Metabolism	RNA processing and modification	mcm5s2U biosynthesis in tRNA	0.1072
RNA Metabolism	RNA processing and modification	mnm5U34 biosynthesis bacteria	0.1072
RNA Metabolism	RNA processing and modification	tRNA modification Archaea	0.1072
RNA Metabolism	RNA processing and modification	tRNA modification Bacteria	0.1072
RNA Metabolism	RNA processing and modification	tRNA modification yeast cytoplasmic	0.1072
Phages, Prophages, Transposable elements, Plasmids	Phages, Prophages	Phage integration and excision	0.1064
Miscellaneous	ZZ gjo need homes	-	0.1064
Amino Acids and Derivatives	Lysine, threonine, methionine, and cysteine	Lysine fermentation	0.1056
Carbohydrates	Fermentation	Acetone Butanol Ethanol Synthesis	0.1056
Carbohydrates	Fermentation	Acetyl-CoA fermentation to Butyrate	0.1056
Carbohydrates	Fermentation	Butanol Biosynthesis	0.1056
Carbohydrates	One-carbon Metabolism	Serine-glyoxylate cycle	0.1056
Clustering-based subsystems	Butyrate metabolism cluster	-	0.1056
Fatty Acids, Lipids, and Isoprenoids	Isoprenoids	Archaeal lipids	0.1056
Fatty Acids, Lipids, and Isoprenoids	Isoprenoids	Isoprenoid Biosynthesis	0.1056
Fatty Acids, Lipids, and Isoprenoids	Polyhydroxybutyrate metabolism	-	0.1056
Metabolism of Aromatic Compounds	Anaerobic degradation of aromatic compounds	Anaerobic benzoate metabolism	0.1056

Top 50 hits were generated by BLASTing sequences to the MG-RAST subsystem database with a minimum alignment length of 50 bp and an E-value cut-off of $1e^{-5}$.

Relative representation in the metagenome was calculated by dividing the number of hits to each category by the total number of hits to all categories.

Table S4.4 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon impacted and non-impacted marine sediment 1 metagenomes.

	Avg. Abundance		
Metabolic Processes	**Non-Impacted sample 1**	**Hydrocarbon-Impacted**	**Contribution %**
Motility and chemotaxis	**0.18**	0.14	11.49
Metabolism of aromatic compounds	0.1	**0.15**	11.48
Photosynthesis	**0.05**	0.02	8.08
Nitrogen metabolism	0.08	**0.11**	7.8
Membrane transport	**0.17**	0.14	5.44

Table S4.5 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon impacted and non-impacted marine sediment 2 metagenomes.

	Avg. Abundance		
Metabolic Processes	**Non-Impacted sample 2**	**Hydrocarbon-Impacted**	**Contribution %**
Metabolism of aromatic compounds	0.11	**0.15**	9.62
Motility and chemotaxis	**0.18**	0.14	9.43
Nitrogen metabolism	0.08	**0.11**	7.82
DNA metabolism	0.18	**0.21**	7.68
Sulfur metabolism	**0.14**	0.12	6.95

Chapter 5

Determining the metabolic footprints of hydrocarbon degradation using multivariate analysis

Submitted as:

Smith RJ, Jeffries TC, Adetutu EM, Fairweather PG, Mitchell JG (2012) The metabolic footprints of hydrocarbon degradation. *PLoS One* (In Review).

5.0 Abstract

The functional dynamics of microbial communities are largely responsible for the clean-up of hydrocarbons in the environment. However, knowledge of the distinguishing functional genes, known as the metabolic footprint, present in hydrocarbon-impacted sites is still scarcely understood. Here, we conducted a multivariate analysis to characterise the metabolic footprints present in hydrocarbon-impacted and non-impacted sediments. Multi-dimensional scaling (MDS) and canonical analysis of principle coordinates (CAP) showed a clear distinction between the two groups. A high relative abundance of genes associated with cofactors, virulence, phages and fatty acids were present in the non-impacted sediments, accounting for 45.7% of the overall dissimilarity. In the hydrocarbon impacted sites, a high relative abundance of genes associated with iron acquisition and metabolism, dormancy and sporulation, motility, metabolism of aromatic compounds and cell signalling were observed, accounting for 22.3% of the overall dissimilarity. These results suggest a major shift in functionality has occurred with pathways more paramount to the degradation of hydrocarbons becoming overrepresented at the expense of other, less essential metabolisms.

5.1 Introduction

Ecosystem functioning is highly dependent on microbial communities (Chapin III *et al*., 1997; Gianoulis *et al*., 2009). These communities are largely defined by biological metabolisms, and are generally thought to be habitat specific (Dinsdale *et al*., 2008b), providing a link between the biology of a given community and the surrounding environment (Gillooly *et al*., 2004). Environmental change can lead to a major shift in the structure and function of the inhabiting microbial consortia (Hemme *et al*., 2010; Kostka *et al*., 2011; Smith *et al*., 2011). Physiological adaptations of microbes have been shown to be highly specific, allowing for the discrimination between chemical stressors (Henriques *et al*., 2007). The identification of defining metabolic pathways of a given ecosystem, known as metabolic footprints, allows for a greater understanding on how the microbial consortia are adapting and responding to environmental change (Gianoulis *et al*., 2009; Röling *et al*., 2010).

Microorganisms are highly responsive to environmental stress, due to a variety of evolutionary adaptions and physiological mechanisms (Schimel *et al*., 2007). The innate ability for microbes to respond and adapt to the world around them means they are often used as biological indicators (Steube *et al*., 2009), and subsequently for bioremediation (Head *et al*., 2006). Many studies have investigated the use of specific microbial taxa as biological indicators (Anderson, 2003; Bonjoch *et al*., 2004; Avidano *et al*., 2005; Mailaa and Cloeteb, 2005), however, previous reports have suggested ecosystems cannot be distinguished by their taxa due to the low variance between habitats (Lozupone and Knight, 2007; Dinsdale *et al*., 2008b; Burke *et al*., 2011). Therefore to gain a comprehensive insight into an ecosystem's

functional response to environmental change, the underlying metabolic footprints need to be elucidated.

Metabolic footprints is a term used to describe an ensemble of biological pathways that typically occur with a combination of environmental variables (Gianoulis *et al*., 2009; Wooley and Ye, 2010). A recent study by Gianoulis *et al*. (2009) used multivariate canonical correlation analysis to describe the metabolic footprints associated with different aquatic environments. These metabolic footprints were thought to arise from differences in evolutionary strategies required to cope with unique environmental variables (Gianoulis *et al*., 2009). Similarly, Dinsdale *et al*. (2008b) used functional differences to discriminate between 9 discrete ecosystems. Here, we employ modern techniques of multivariate analysis with few assumptions to determine the metabolic footprints of hydrocarbon-impacted environments.

The long-lasting toxicity of xenobiotics makes their metabolism by microbial communities widely studied (Singleton, 1994). Petroleum hydrocarbons are a common target for bioremediation because they are widespread and persistent (Röling *et al*., 2002; Vinas *et al*., 2005; Chikere *et al*., 2011; Kostka *et al*., 2011; Liang *et al*., 2011). While the optimal taxa and environmental conditions for optimal degradation of hydrocarbons are well established (Xu *et al*., 2003; Walworth *et al*., 2007; Yakimov *et al*., 2007; Singh *et al*., 2011), the effectiveness of a natural community to bioremediate is less well understood (Chakraborty *et al*., 2012).

Advances in metagenomic technologies have allowed for the direct sequencing of environmental microbial communities (Kennedy *et al*., 2010), greatly increasing our potential to understand the metabolic processes being undertaken by the indigenous microbial communities. A recent study by Yergeau *et al*. (2012) used metagenomic

sequencing technologies to characterise the structure and function of an active soil microbial community in a hydrocarbon contaminated Arctic region. However, this study primarily focused on the taxa present, and not the defining metabolic activities associated with hydrocarbon contamination. Thus, knowledge on the distinguishing functional genes present in hydrocarbon contaminated environments is still lacking. The aim of the present study was to compare hydrocarbon-impacted sites to non-impacted sites, and provide insight into the key metabolic functions present following hydrocarbon impact, thus elucidating the metabolic footprints for hydrocarbon contamination.

5.2 Materials and Methods

5.2.1 Data Collection

To determine the functionality of microbial communities inhabiting hydrocarbon-impacted and non-impacted environments, publicly available datasets were chosen from the MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) pipeline version 3.0 (Meyer *et al*., 2008). Due to constraints in the database, a total of 4 datasets were used to represent hydrocarbon-impacted environments, while 5 datasets were used for non-impacted environments (Table S5.1). BLASTX was performed on all datasets, with a minimum alignments length of 50 bp and an E-value cut-off of $E<1e^{-5}$ (Dinsdale *et al*., 2008b), to identify hits to the subsystems database.

5.2.2 Data Analysis

To statistically investigate the differences between metagenomes from hydrocarbon-impacted sites to metagenomes from un-impacted sites, heatmaps were generated containing the relative proportion of hits to the subsystem database in MG-RAST.

Heatmaps had been standardized and scaled to account for differences in sequencing effort and read lengths. Statistical analysis was conducted on square-root transformed data to reduce the impact of dominant metabolisms using the software package Primer 6 for Windows (Version 6.1.13, Primer-E, Plymouth) (Clarke and Gorley, 2006). Level 1 hierarchial classification was used to determine the overall differences in metabolic potential (Dinsdale *et al*., 2008b; Gianoulis *et al*., 2009).

Differences in metabolic potential between hydrocarbon impacted and non-impacted sediments were analysed using the PERMANOVA+ version 1.0.3 3 add-on to PRIMER (Anderson and Robinson, 2001; Anderson *et al*., 2008). Non-metric Multi-Dimensional scaling (MDS) of Bray-Curtis similarities was performed as an unconstrained ordination method to graphically visualise multivariate patterns in the metabolic processes associated hydrocarbon-impacted and non-impacted sediment metagenomes. Metagenomes were further analysed using canonical analysis of principle coordinates (CAP) on the sum of squared canonical correlations as a constrained method, to determine if there was any significant trend between metabolic processes according to hydrocarbon impact. The *a priori* hypothesis that the metabolisms between the two groups were different was tested in CAP (Anderson *et al*., 2008) by obtaining a *P*-value using 9999 permutations.

Where significant differences were found using CAP, the percent contribution of each metabolism to the separation between the hydrocarbon-impacted and non-impacted sediments were assessed using similarity percentage (SIMPER) analysis (Clarke, 1993). The resulting top 90 percent of all metabolisms were used to determine the shifts in metabolic potential between the groups. To determine those metabolisms that were consistently contributing to the overall dissimilarity between

the hydrocarbon-impacted and non-impacted groups, the ratio of the average dissimilarity to standard deviation (Diss/SD) was used. A Diss/SD ratio of greater than 1.4 was used to indicate key discriminating metabolisms (Clarke and Warwick, 2001).

5.3 Results

MDS analysis revealed a clear separation of data between the hydrocarbon-impacted and non-impacted sediment metagenomes (Fig. 5.1). CAP analysis confirmed this separation showing significant differences between the two groups (P = 0.008). A strong association between the multivariate data and the hypothesis of metabolic difference was indicated by the large size of their canonical correlations (δ^2 = 0.83). The first canonical axis (m = 1) was used to separate the samples (Fig. 5.2). Cross validation of the CAP model showed all samples were correctly classified to hydrocarbon-impacted and non-impacted sediments, hence with a zero mis-classification rate (Table 5.1).

SIMPER analysis revealed the main metabolic processes contributing to the dissimilarity in the non-impacted sediments when compared to the hydrocarbon-impacted sediments, were genes associated with cofactors, virulence, phages and fatty acids, together accounting for 45.71% of the overall dissimilarity. Genes associated with protein metabolism, carbohydrates, amino acids, clustering-based subsystems, potassium metabolism, respiration, RNA metabolism, nucleosides and cell wall were also higher in the non-impacted site compared to the impacted sites, collectively contributing to 9.88% of the overall dissimilarity (Table 5.2 and S5.2).

Conversely, the main metabolic processes associated with the hydrocarbon impacted sediments were iron acquisition and metabolism, dormancy and sporulation, motility,

metabolism of aromatic compounds and cell signalling accounting for 22.3% of the overall dissimilarity between the two groups (Table 5.2). Genes associated with nitrogen, phosphorus and sulfur metabolisms were also higher in the hydrocarbon impacted site, collectively accounting for 2.5% of the dissimilarity to the non-impacted sites. Regardless of percent contribution, however, all metabolic processes, with the exception of secondary metabolism and photosynthesis, are likely good discriminators for hydrocarbon-impacted or non-impacted sediments, indicated by a dissimilarity/standard deviation ratio (Diss/SD) of greater than 1.4 (Clarke and Warwick, 2001) (Table 5.2 and S5.2).

5.4 Discussion

Microbial communities are known to respond to hydrocarbon contamination at the genotypic level (Langworthy *et al*., 1998; Siciliano *et al*., 2003; Head *et al*., 2006). Thus, a major goal in the study of bioremediation is to identify the key metabolic processes being undertaken by the inhabiting microbial communities (Watanabe, 2001; Chakraborty *et al*., 2012). Here, we report the first metagenomic study to identify the overall metabolic footprints associated with discriminating hydrocarbon-impacted versus non-impacted sediment samples.

Unconstrained (MDS) and constrained (CAP) multivariate analyses showed a significant difference (P = 0.008; Table 5.1) between the relative abundances of metabolisms for hydrocarbon-impacted and non-impacted sediment (Fig. 5.1 and 5.2). The similarities between constrained and unconstrained ordinations likely reflect the single hydrocarbon impact pressure. This is supported by the CAP analysis, which shows that the majority of the variance is expressed on just the first canonical axis, with a squared canonical correlation (δ^2) of 0.83 (Table 5.1). A

recent hydrocarbon-based study used high throughput functional gene array technology to show that all microbial samples with hydrocarbon contamination grouped together indicative of similar functional patterns (Liang *et al*., 2011). Furthermore, it has been shown that differences in metabolic processes could be used to predict the biogeochemical status of the environment (Dinsdale *et al*., 2008b). Thus, the clear separation between data points in the MDS and CAP plots indicates the hydrocarbon-impacted sediment samples can be readily distinguished based on metabolic processes.

The majority of the separation between the two groups was explained by a higher relative abundance of genes associated with cofactors, virulence, phages and fatty acids, collectively accounting for 45.71% of the dissimilarity in the non-impacted sediment samples when compared to the impacted sites (Table 5.2). Those microbes capable of surviving following hydrocarbon impact become dominant, leading to a major shift in the structure of the community (Vinas *et al*., 2005; Wu *et al*., 2008). This shift in structure is generally coupled with the reduction of non-essential metabolic pathways (Liang *et al*., 2009; Hemme *et al*., 2010). Thus, the high degree of dissimilarity driven by the non-impacted sediments, suggests the major factor causing the differences between the two groups can be explained by a shift in functionality, which has led to the reduction in non-essential metabolisms following hydrocarbon impact.

The reduction in non-essential metabolic pathways was coupled with a subsequent increase in pathways associated iron acquisition and metabolism, dormancy and sporulation, motility, metabolism of aromatic compounds and cell signalling (Table 5.2). These pathways have all previously been linked to stressed environments (Ford,

2000; Schneiker *et al*., 2006; Suenaga *et al*., 2007; Hemme *et al*., 2010), suggesting the microbial communities inhabiting the hydrocarbon-impacted environments are exerting more energy on pathways essential to the utilization of carbon and survival.

The degradation of hydrocarbons is often hindered by the requirement to come into direct contact with hydrocarbon substrates (Ron and Rosenberg, 2002). Therefore, many microorganisms capable of catabolising hydrocarbons have shown chemotaxis abilities allowing them to move towards, and subsequently degrade the contaminant at a higher rate (Ortega-Calvo *et al*., 2003; Peng *et al*., 2008; Fernández-Luqueño *et al*., 2011). This degradation ability is then often further enhanced by the secretion of biosurfactants, which increase the availability of hydrocarbons in the soil (Venkata Mohan *et al*., 2006). Thus, the increase in motility and chemotaxis genes suggest the microbial communities are increasing metabolisms that will allow for direct contact with hydrocarbon compounds (Table 5.2).

Following direct contact, the microbial communities must have genes that allow for the catabolism of hydrocarbons. Petroleum hydrocarbons are comprised of a complex mixture of compounds including cycloalkanes, alkanes, polycyclic aromatic hydrocarbons, aromatics and phenolics (Hamamura *et al*., 2006). Previous studies have shown an increase in genes associated with the breakdown of these compounds in hydrocarbon contaminated environments (Yergeau *et al*., 2009; Liang *et al*., 2011). Thus, a higher relative abundance of metabolism of aromatic compound genes in the hydrocarbon-impacted sediments when compared to the non-impacted sediments is consistent with a community optimising its ability to utilise hydrocarbon as an energy source (Table 5.2).

Table 5.1 Results of CAP analysis for metabolisms associated with hydrocarbon impacted and non-impacted sediment metagenomes

Group	Allocation success (%)	δ^2	*P*-value
Hydrocarbon-impacted sediments	100	0.829	0.008
Non-impacted sediments	100	0.829	0.008

Following hydrocarbon contamination, microbial communities must adapt to tackle the sudden increase in carbon availability and subsequent loss of limiting nutrients such as nitrogen and phosphorus and in some cases iron (Beller *et al*., 1992; Head *et al*., 2006; Schneiker *et al*., 2006). As a result, an increase in genes associated with nitrogen, phosphorus and iron metabolism have been shown, allowing for effective scavenging mechanisms (Smith *et al*., unpublished data). Hydrocarbon impact has also been shown to stimulate the sulfur cycle significantly, indicating its importance when dealing with crude oil contamination (Kleikemper *et al*., 2002). Our results indicate there has been an increase in nitrogen, phosphorus, sulfur and iron metabolites in the hydrocarbon-impacted sediments when compared to non-impacted sediments. Furthermore, genes associated with cofactors, amino acid pathways, carbohydrates and protein metabolisms were all reduced in the hydrocarbon-impacted sites (Table 5.2 and S5.2). Taken together, these results suggest the microbial communities are expending most of their energy scavenging key nutrients needed for bioremediation of hydrocarbons, leading to the subsequent decrease in pathways associated with more complex carbohydrate and protein metabolisms and growth.

Although some pathways contributed to the dissimilarity between the two groups more than others, all metabolisms with the exception of secondary metabolism and photosynthesis were identified as being consistent distinguishing metabolisms (Table 5.2 and S5.2). This suggests all are metabolic footprints of their given environment, indicating the overall metabolic signature is different between groups. In nature, microbial communities are typically composed of mixed communities characterised by an intricate network of metabolic processes (Pelz *et al*., 1999). Consequently, our

results indicate a complete overview of the metabolites present within the inhabiting microbial consortia is needed to effectively characterise an environment.

5.5 Conclusion

Our data indicates the hydrocarbon-impacted sediment samples can be distinguished from non-impacted sediments based on their metabolic signatures. These signatures include metabolisms associated with iron acquisition and metabolism, dormancy and sporulation, motility, metabolism of aromatic compounds, cell signalling and nitrogen, phosphorus and sulfur metabolism. Our data also indicated that the majority of the dissimilarity, however, was due to a reduction of functional genes associated with cofactors, virulence, phages and fatty acids. This study elucidated the intricate network of functional genes associated with hydrocarbon impact, allowing for the characterisation of metabolic footprints.

5.6 Acknowledgements

The authors gratefully acknowledge the funding provided by the Australian Research Council. R. J. Smith is the recipient of a Flinders University Research Scholarship (FURS).

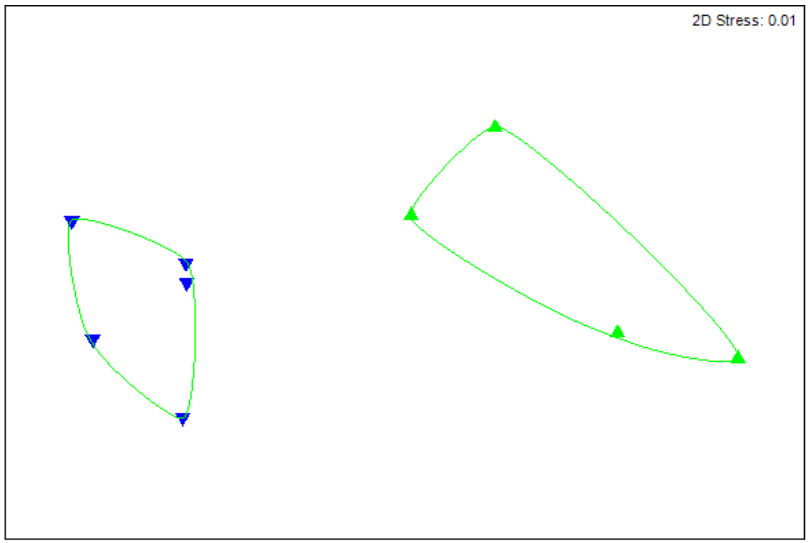

Figure 5.1 Comparison of hydrocarbon-impacted sediments (green) and non-impacted sediments (blue). MDS profile is derived from a Bray-Curtis similarity matrix calculated from the square-root transformed abundance of DNA fragments matching the subsystems database, level hierarchial system 1 (BLASTX E-value $<1e^{-5}$). The light green polygons depict significantly different groupings ($P < 0.05$) as calculated by similarity profile (SIMPROF) analysis.

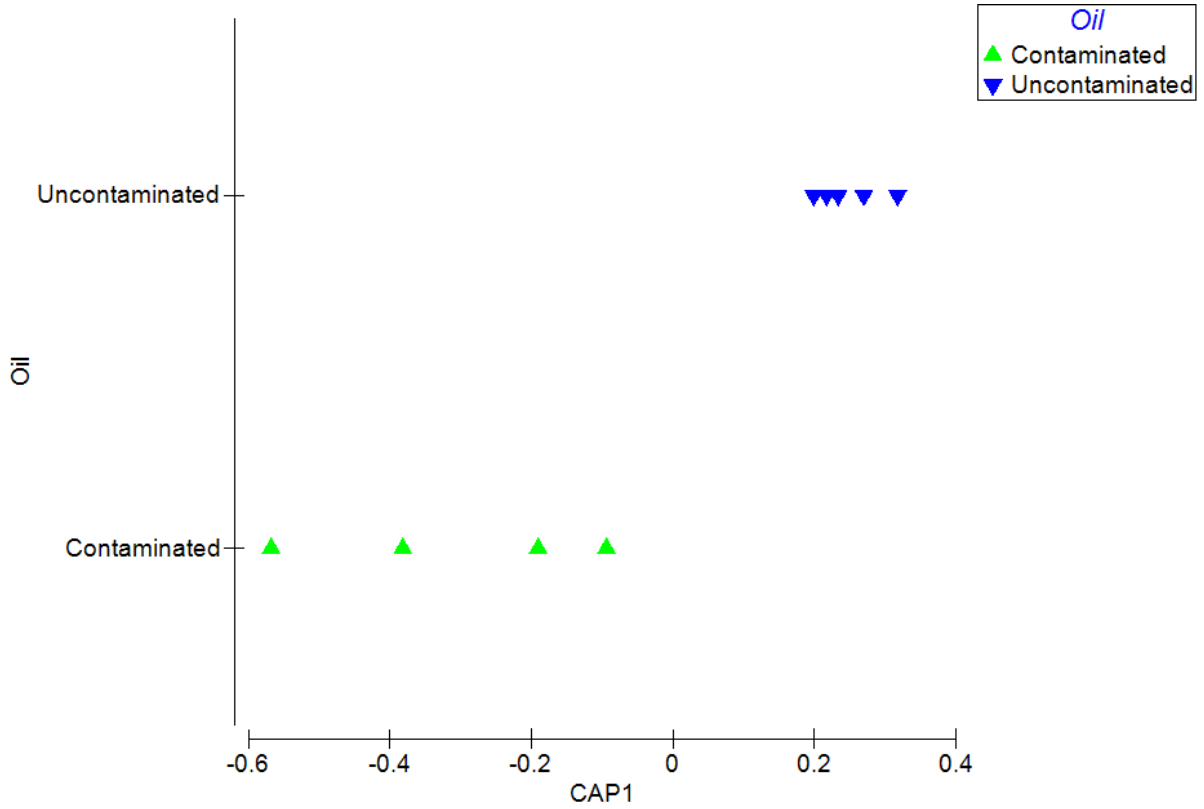

Figure 5.2 Comparison of hydrocarbon-impacted sediments (green) and non-impacted sediments (blue). CAP analysis is derived from the sum of squared correlations of DNA fragments matching the subsystems database, level hierarchial system 1 (BLASTX E-value $<1e^{-5}$).

Table 5.2 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon-impacted and non-impacted sediment metagenomes. Average dissimilarity between the two groups is 1.78%. Only metabolisms that were consistent (i.e. Diss/SD > 1.4) are shown here. The larger value in each case (i.e. the potential indicator of that condition) is shown in bold.

	Avg. Abundance			
Metabolic Processes	**Hydrocarbon-Impacted**	**Non-Impacted**	**Diss/ SD**	**Cum %**
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.1	**0.19**	2.24	11.43
Virulence, Disease and Defence	0.1	**0.19**	2.24	22.86
Phages, Prophages, Transposable elements, Plasmids	0.1	**0.19**	2.24	34.29
Fatty Acids, Lipids, and Isoprenoids	0.1	**0.19**	2.24	45.71
Iron acquisition and metabolism	**0.84**	0.79	1.63	52.68
Dormancy and Sporulation	**0.71**	0.68	1.49	57.48
Motility and Chemotaxis	**0.83**	0.81	1.58	61.17
Metabolism of Aromatic Compounds	**0.87**	0.85	1.73	64.81
Secondary Metabolism	**0.76**	0.75	1.16	68.32
Regulation and Cell signalling	**0.86**	0.83	1.86	71.55
Protein Metabolism	0.94	**0.96**	3.42	74.53
Carbohydrates	0.97	**1**	3.5	77.49
Nitrogen Metabolism	**0.84**	0.82	1.74	80.17
Photosynthesis	0.69	0.69	1.3	82.75
Amino Acids and Derivatives	0.96	**0.98**	2.89	85.24
Clustering-based subsystems	0.98	**0.99**	1.96	87.06
Miscellaneous	0.94	**0.96**	3.14	88.7

Cut-off percentage = 90%, Diss=dissimilarity; SD=Standard Deviation; Cum %=cumulative percentage of contribution to overall dissimilarity, Avg. Abundance values are reported for square-root transformed data

Table S5.1 Summary of publicly available metagenomes used in this study.

MG-RAST ID	Description/Reference
4453082.3	Hydrocarbon contaminated foreshore
4453072.3	Hydrocarbon contaminated biopile
4449126.3	Biopiles 2006 (Yergeau *et al*., 2012)
4450729.3	Biopile 2005 (Yergeau *et al*., 2012)
4446341.3	Marine sediment 1 (Jeffries *et al*., 2011a)
4446342.3	Marine sediment 2 (Jeffries *et al*., 2011a)
4440984.3	Coorong sediment 1 (Jeffries *et al*., 2011a)
4441020.3	Coorong sediment 2 (Jeffries *et al*., 2011a)
4441021.3	Coorong sediment 3 (Jeffries *et al*., 2011a)

Table S5.2 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon-impacted and non-impacted sediment metagenomes. Shows all metabolisms, including inconsistent ones (i.e. Diss/SD < 1.4). Average dissimilarity between the two groups is 1.78 %. Bold values show either the condition with the higher average abundance (i.e. a potential indicator of that condition) or Diss/SD ratios that are consistent (i.e. > 1.4).

Metabolic Processes	Avg. Abundance Hydrocarbon-Impacted	Avg. Abundance Non-Impacted	Diss/SD	Cum %
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.1	**0.19**	**2.24**	11.43
Virulence, Disease and Defence	0.1	**0.19**	**2.24**	22.86
Phages, Prophages, Transposable elements, Plasmids	0.1	**0.19**	**2.24**	34.29
Fatty Acids, Lipids, and Isoprenoids	0.1	**0.19**	**2.24**	45.71
Iron acquisition and metabolism	**0.84**	0.79	**1.63**	52.68
Dormancy and Sporulation	**0.71**	0.68	**1.49**	57.48
Motility and Chemotaxis	**0.83**	0.81	**1.58**	61.17
Metabolism of Aromatic Compounds	**0.87**	0.85	**1.73**	64.81
Secondary Metabolism	0.76	0.75	1.16	68.32
Regulation and Cell signalling	**0.86**	0.83	**1.86**	71.55
Protein Metabolism	0.94	**0.96**	**3.42**	74.53
Carbohydrates	0.97	**1**	**3.5**	77.49
Nitrogen Metabolism	**0.84**	0.82	**1.74**	80.17
Photosynthesis	0.69	0.69	1.3	82.75
Amino Acids and Derivatives	0.96	**0.98**	**2.89**	85.24
Clustering-based subsystems	0.98	**0.99**	**1.96**	87.06
Miscellaneous	0.94	**0.96**	**3.14**	88.7
Potassium metabolism	0.79	**0.8**	**1.45**	90.27
Respiration	0.89	**0.9**	**1.51**	91.79
Phosphorus Metabolism	**0.84**	0.83	**1.41**	93.3
RNA Metabolism	0.92	**0.93**	**1.83**	94.62
Sulfur Metabolism	**0.84**	0.83	**1.6**	95.89
Nucleosides and Nucleotides	0.88	**0.89**	**1.58**	97.03
Cell Wall and Capsule	0.91	**0.92**	**1.62**	97.74
Stress Response	0.89	0.89	**1.43**	98.38
Cell Division and Cell Cycle	0.84	0.84	1.39	98.99
DNA Metabolism	0.91	0.91	1.24	99.54
Membrane Transport	0.9	0.9	1.28	100

Diss=dissimilarity; SD=Standard Deviation; Cum %=cumulative percentage of contribution to overall dissimilarity, Avg. Abundance values are reported for square-root transformed data

Chapter 6

Towards elucidating the metagenomic signatures for impacted environments

6.0 Abstract

Anthropogenic modification has led to the accumulation of toxic xenobiotics worldwide. Due to their resilience to environmental change, microbial communities are increasingly used as indicator organisms to monitor polluted sites. The enormous abundance and diversity of microbial communities, however, has often hindered our ability to characterise polluted sites based on their microbial communities. Here, we employed a constrained multivariate analysis, canonical analysis of principal coordinates (CAP), to generate metagenomic signatures for three common forms of environmental impacts; agricultural effluent, hydrocarbon and wastewater. Significant differences between impacted environments were shown, with a 75% and 100% allocation success for hydrocarbon and agriculturally impacted sites, respectively, however, wastewater could not be consistently distinguished. The main distinguishing metabolic processes associated with agricultural-impacted environments were genes associated with cofactors, virulence, phages and fatty acids. Conversely, the main distinguishing genes associated with hydrocarbon-impacted sites were iron acquisition and metabolism, photosynthesis, aromatic compound degradation, dormancy and motility. Taken together, these results indicate that a markedly different response by the microbial communities to contaminant type.

6.1 Introduction

Microbial communities typically consist of mixed consortia, which are characterised by intricate networks of metabolic and phylogenetic diversity (Pelz *et al*., 1999). These complex networks allow for innate flexibility, whereby the microbial communities are able to adapt swiftly to environmental change, including the introduction of xenobiotic contamination (Marzorati *et al*., 2008). Furthermore, the biodiversity within a microbial community generally leads to a high degree of resilience and biological functionality (Griffiths *et al*., 2001; Loreau *et al*., 2001). This rapid response to the changing world, as well as their inherent survival mechanisms, means that microbial communities are often used as biological indicators, or signatures, for a given environment (Dinsdale *et al*., 2008b; Gianoulis *et al*., 2009; Steube *et al*., 2009).

Shifts in microbial community composition whereby rare taxa or metabolic processes become more prominent are often linked to environmental change (Sogin *et al*., 2006; Dinsdale *et al*., 2008b; Jeffries *et al*., 2011a; Jeffries *et al*., 2011b; Smith *et al*., 2011). Furthermore, previous studies have shown that microbial communities often respond at a genotypic level before any disturbance is seen at the taxonomic level (Parnell *et al*., 2009). Due to this genotypic response, it is suggested that ecosystems are better described by their metabolic potential rather than by their taxa (Lozupone and Knight, 2007; Burke *et al*., 2011). However, whether there is a loss of information between the different levels of taxonomic and metabolic resolution is yet to be determined.

Advances in high-throughput sequencing technologies have allowed for a greater sensitivity when generating microbial profiles of environmental systems (Kennedy *et*

al., 2010; Xing *et al*., 2012). The result is a greater understanding of the abundance and distribution of taxa and genes that establish as a result of environmental change. The distinguishing taxa and metabolic potential of an environment responding to environmental impact can then be used to generate metagenomic signatures.

Many studies have used multivariate analysis to identify distinguishing characteristics in the microbial communities inhabiting different environmental systems (Buyer and Drinkwater, 1997; Hernesmaa *et al*., 2005; Dinsdale *et al*., 2008a; Gianoulis *et al*., 2009; Liang *et al*., 2011). The majority of these studies used constrained ordinations such as canonical discriminant analysis (CDA) and principal component analysis (PCA) (Buyer and Drinkwater, 1997; Hernesmaa *et al*., 2005; Dinsdale *et al*., 2008b; Liang *et al*., 2011). However, these methods are restricted in that PCA cannot be performed on a dataset containing more observations (samples) than variables (taxa/metabolic processes), and CDA should be performed on a dataset where there are at least three times as many observations than variables (Williams and Titus, 1988; Buyer and Drinkwater, 1997). This results in the need to reduce the number of variables prior to analysis (Buyer and Drinkwater, 1997). Microbial communities, however, comprise intricate networks whereby a large number of individuals/metabolic processes are important in the overall ecosystems functioning (Pelz *et al*., 1999). Thus, the community as a whole should be considered when categorising a given environment (Smith *et al*., unpublished data).

Canonical analysis of principal coordinates (CAP) is also a constrained multivariate analysis, however, unlike CDA and PCA it allows for the characterisation of whole communities as it is not limited by observation size (Anderson and Willis, 2003). This multivariate analysis has been used in many studies to determine how microbial

communities respond to various environmental conditions (Bastias *et al*., 2006; Cookson *et al*., 2007; Baker *et al*., 2009; Lear and Lewis, 2009); however, to date, it has not been employed to generate metagenomic signatures for various impacted environments. Thus, we sought to construct a taxonomic and metabolic profile of microbial communities responding to various forms of environmental impacts, in order to generate metagenomic signatures using CAP. The information generated from this study can then be used to determine the biological indicators for xenobiotic pollution as well as to better understand the role microbes play in the catabolism of toxic compounds.

6.2 Materials and Methods

6.2.1 Data Collection

To statistically investigate the metagenomic signatures for three common forms of environmental impacts; agriculture, hydrocarbon and wastewater (Table S6.1), heatmaps were generated in MetaGenomics Rapid Annotation using Subsystem Technology (MG-RAST) pipeline version 3.0 (Meyer *et al*., 2008), which had been standardized and scaled to account for differences in sequencing effort and read lengths. Taxonomic profiles were generated using the normalized abundances of sequences matches to the SEED database (Overbeek *et al*., 2005), while metabolic profiles were generated successively using the normalized abundances of sequences matches to the subsystems database. An E-value cut-off of $E<1e^{-5}$ and a minimum alignment length of 50 bp was used to identify hits. Heatmaps were generated using the phylum, class, order, family and genus levels of resolution available in MG-RAST for taxonomy and hierarchial level 1 and 2 for metabolism. Statistical analyses were conducted on square-root transformed data using the statistical software

package Primer 6 for Windows (Version 6.1.13, Primer-E, Plymouth) (Clarke and Gorley, 2006).

6.2.2 Data Analysis

To determine whether there was any loss of information between the level of resolution for taxonomy and metabolism, the program RELATE in the Primer package was used to calculate the rank correlation between each pair of classifications (Clarke, 1993). Differences in the overall taxonomy and metabolic potential between the impacted environments were analysed using PERMANOVA+ version 1.0.3 3 (Anderson *et al*., 2008). The CAP on the sum of squared canonical correlations (Anderson and Robinson, 2001) was performed to graphically illustrate the multivariate patterns associated with the impacted environments for taxonomy and metabolism. Significant trends between the overall taxonomy and metabolic processes at each site were determined using the sum of squared canonical correlations. The *a priori* hypothesis that either the taxonomy or metabolisms between the two groups were different was tested using 9999 permutations. Based on RELATE results, CAP ordinations were generated using phylum and hierarchy level 1 for taxonomy and metabolism, respectively.

Where statistically significant differences were shown using CAP analysis, similarity percentage (SIMPER) analysis (Clarke, 1993) was conducted to determine the main taxa and metabolisms driving the dissimilarity between contamination types. The average dissimilarity to standard deviation (Diss/SD) ratio was used to determine the taxa and metabolisms that were consistently contributing to the overall dissimilarity between types, whereby key discriminating taxa and metabolisms were indicated by a Diss/SD ratio of at least 1.4 (Clarke and Warwick, 2001).

Table 6.1 Spearman rank correlation coefficients for comparisons of similarity matrices for each pair of taxonomic and metabolic level of resolution. All correlations were significant at $P < 0.001$.

Taxonomy		**Genus**	**Family**	**Order**	**Class**
	Phylum	0.713	0.785	0.847	0.908
	Class	0.736	0.823	0.939	-
	Order	0.816	0.89	-	-
	Family	0.944	-	-	-
Metabolism		**Level 2**			
	Level 1	0.773			

6.3 Results

A reduction in the rank coefficients between the different levels of resolution for taxonomy and metabolism was seen, with a higher rank coefficient of 0.9 for comparisons between phylum and class level compared to 0.7 for comparisons between phylum and genus level and hierarchial level 1 and 2 (Table 6.1). Closer ranks, family/genus or phylum/class, had higher correlations than more distant pairs, family/phylum or genus/class. However, all combinations of taxonomic and metabolic resolution were significantly correlated ($P < 0.001$) indicating similar results were seen irrespective of hierarchial classification (Table 6.1). Thus, to create a robust set of metagenomic signatures, all further analyses were conducted on phylum level and hierarchial level 1 for taxonomy and metabolism, respectively. When comparing metabolism to taxonomy, there was no significant correlation between phylum level and hierarchial level 1 ($P = 0.09$) indicating the information gained from taxonomy and metabolic potential differs.

CAP ordination revealed a clear separation of data between the impacted environments impacted environments based on either taxonomy or metabolic potential (Fig. 6.1 and 6.2); however only the metabolic potential showed significant differences between the environmental contaminants ($P = 0.008$) (Table 6.2), thus the remainder of this manuscript will focus on the differences in metabolic potential. A strong association was seen between the multivariate data and the hypothesis of metabolic differences, indicated by the large size of their canonical correlations (hierarchial level 1: $\delta^2 = 0.86$). Cross validation of the CAP model showed 75% of samples overall were correctly classified to their impacted environments. More specifically, 75% and 100% of hydrocarbon and agricultural impacted sites,

respectively, were correctly allocated, while only 50% and 0% of wastewater and pristine sites were correctly classified (Table 6.2).

Based on CAP ordinations as well as allocation success percentages, SIMPER analysis was used to determine distinguishing metabolic processes for the oil and agricultural impacted sites only. SIMPER analysis revealed the main metabolic processes contributing to the dissimilarity in the agricultural impacted environments when compared to the hydrocarbon impacted environments were genes associated with cofactors, virulence, phages and fatty acids, collectively accounting for 48% of the overall dissimilarity between these two types. Genes associated with protein metabolism, carbohydrates, amino acids and clustering based subsystems were also higher in the agricultural impacted sites when compared to hydrocarbon impacted sites, collectively contributing to another 18.4% of the overall dissimilarity (Table 6.3 and S6.2).

Alternatively, the main metabolic processes associated with hydrocarbon impact were genes related to iron acquisition and metabolism, photosynthesis, aromatic compound degradation, dormancy and motility, collectively contributing to 20.1% of the overall dissimilarity (Table 6.3 and S6.2). Genes associated with regulation and nitrogen metabolism were also higher in the hydrocarbon impacted sites when compared to agricultural impacted sites, collectively accounting for 5.2% (Table 6.3 and S6.2). Furthermore, all metabolic processes, with the exception of potassium metabolism, secondary metabolism and cell division were consistently distinguishable between agricultural and oil impacted environments, indicated by a dissimilarity/standard deviation ration (Diss/SD) of greater than 1.4 (Clarke and Warwick, 2001).

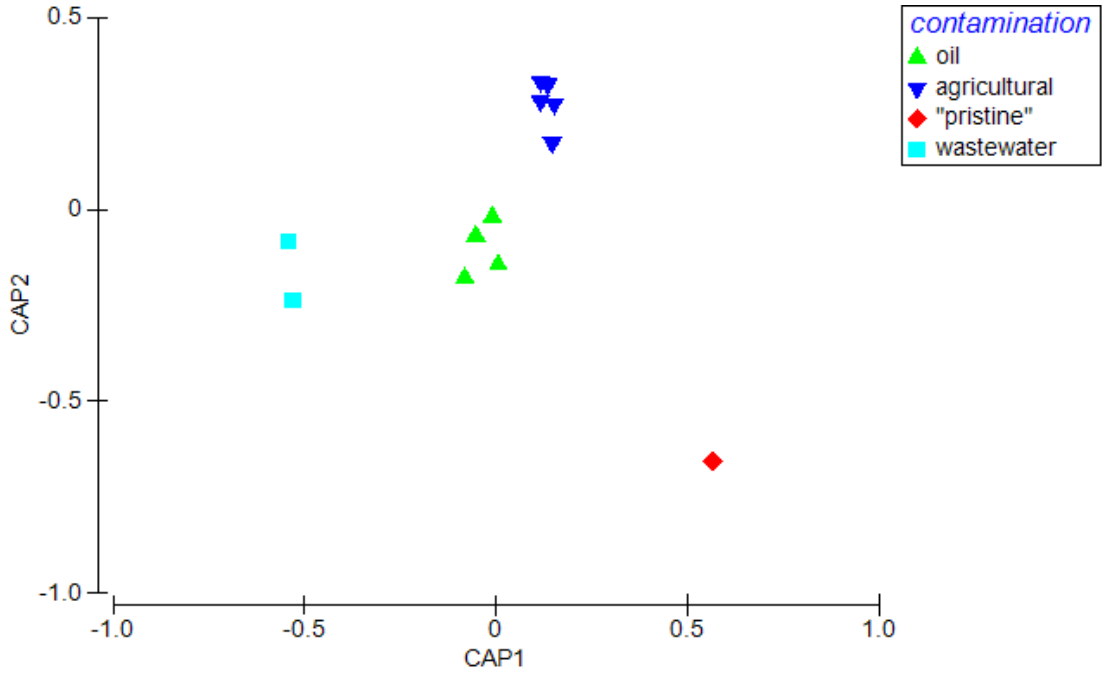

Figure 6.1 Taxonomic comparison of impacted environments. CAP analysis is derived from the sum of squared correlations of DNA fragments matching the SEED database, phylum level (BLASTX E-value $<1e^{-5}$).

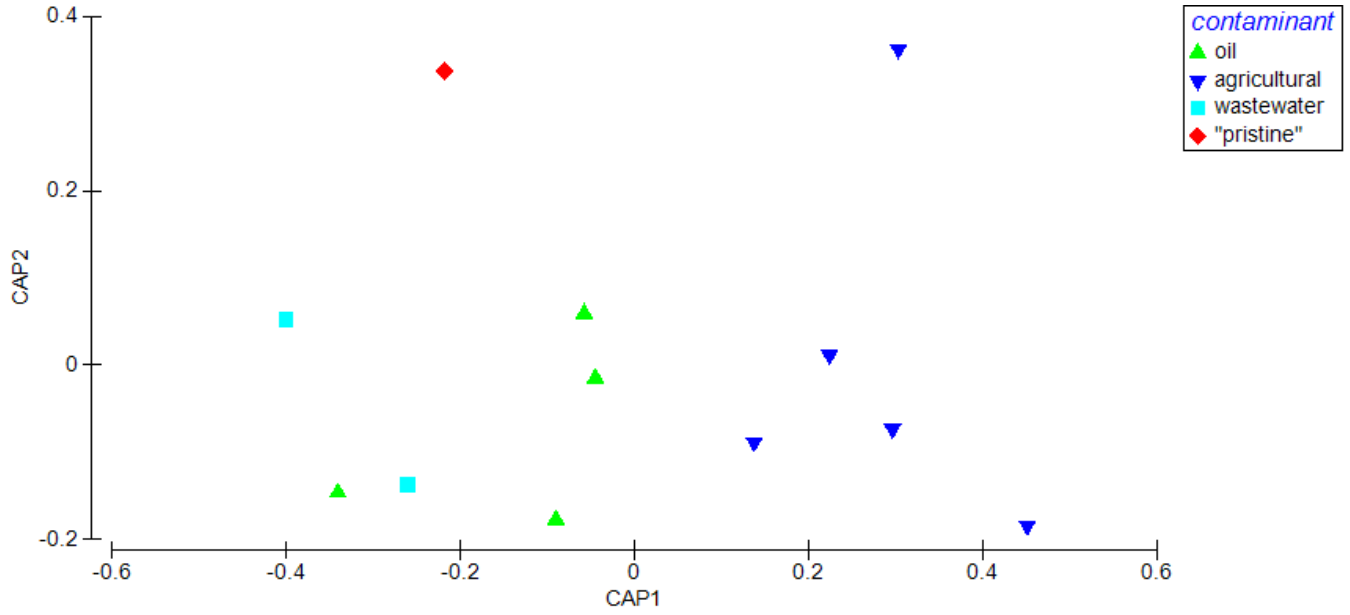

Figure 6.2 Metabolic comparison impacted environments. CAP analysis is derived from the sum of squared correlations of DNA fragments matching the subsystems database, level hierarchial system 1 (BLASTX E-value $<1e^{-5}$).

6.4 Discussion

Anthropogenic pollution has led to the accumulation of a wide variety of toxic xenobiotics causing detrimental effects to pristine ecosystems worldwide (Naeem and Li, 1997). Understanding the intimate relationship between environmental anthropogenic disturbances and shifts in microbial communities is now recognised as an imperative ecological parameter in monitoring polluted sites (Gelsomino *et al*., 2006). Here, we sought to distinguish between various contaminant types by the inhabiting microbial communities, in order to generate metagenomic signatures for polluted environments.

RELATE analysis showed a significant correlation ($P < 0.001$) between all levels of taxonomic and metabolic hierarchy (Table 6.1), indicating there is no significant loss of information between the different levels of resolution. This result is consistent with previous studies that have shown changes to environmental conditions caused by anthropogenic disturbances have led to major shifts in microbial community structure and functionality that become evident across multiple levels of resolution (Hemme *et al*., 2010; Jeffries *et al*., 2011a; Smith *et al*., 2011).

Alternatively, there was a low level of correlation when comparing structure to function suggesting that extra information can be gained from one over the other. It is generally thought that species diversity determines community stability, whereby a higher diversity correlates to a higher inherent stability (Naeem and Li, 1997). However, more recently, studies have shown that even those communities with low species diversity are still able to maintain a degree of plasticity through a high genotypic diversity within key species (Bailey *et al*., 2006; Crutsinger *et al*., 2006). Moreover, when stable/species-rich environments are disturbed, a reduction in

genotypic diversity has been shown to occur regardless of species diversity maintenance (Parnell *et al*., 2009). Therefore, the low level of correlation between structure and function is likely driven by an incomplete story generated from taxonomy alone.

CAP analysis showed a significant difference (P = 0.008; Table 6.2) between the relative abundances of metabolisms for impacted environments (Fig. 6.2). In particular, hydrocarbon and agricultural impacted environments were found to have the highest allocation success, 75% and 100% respectively, when compared to wastewater and pristine sites, 50% and 0%, respectively (Table 6.2). The higher misclassification rate for wastewater and pristine sites, when compared to hydrocarbon and agricultural impacted sites was likely driven by the larger sample size for hydrocarbon and agricultural environments than for the wastewater and pristine environments. Previous studies have shown the ability to measure the impact of pollution through molecular fingerprinting and signature biomarkers (White *et al*., 1998). Furthermore, measures of functional stability, in particular resistance genes, have proven to be useful in distinguishing between various environmental impacts in soil (Griffiths *et al*., 2001). Thus, CAP analysis suggests the impacted environments have acquired microbial communities with differing metabolic functions, which have allowed for our ability to distinguish between contaminant types.

SIMPER analysis revealed the main distinguishing metabolic processes associated with agricultural impacted environments were genes associated with cofactors, virulence, phages, fatty acids, protein metabolism, carbohydrates, amino acids and clustering based subsystems (Table 6.3 and S6.2), collectively accounting for 66.4% of the overall dissimilarity to the hydrocarbon-impacted environments. A recent

metagenomic study showed a relatively high proportion of viral sequences, 9%, in groundwater affected by agricultural impact (Smith *et al*., 2011). Furthermore, a study by Dinsdale *et al*. (2008a) showed a higher proportion of pathogens in human-impacted when compared to non-impacted marine environments. Therefore, the higher proportion of virulence and phage genes in the agricultural impacted environments when compared to the hydrocarbon-impacted environments is consistent with reports that human-impact, or more specifically agricultural impact, can lead to an increase in overall viral numbers.

Agricultural practices are known to increase the deposition of nutrients into the surrounding environment (Haberl *et al*., 2007; Barnosky *et al*., 2012). Previous studies have shown that an increase of nutrients via agricultural impact can lead to an increase in microbial productivity (Smith *et al*., 2011). Alternatively, hydrocarbon impact has been shown to lead to a reduction in genotypic diversity, whereby only the essential metabolisms remain (Hemme *et al*., 2010; Liang *et al*., 2011). This is thought to be due to the toxic effect of hydrocarbon pollution which in turn can lead to a community exerting more energy on survival than on growth and productivity (Delille and Delille, 2000; Smith *et al*., unpublished data). Thus, an increase in genes associated with protein metabolism in the agricultural impacted environments (Table 6.3) is consistent with a more active community when compared to the hydrocarbon impacted environments (Urich *et al*., 2008).

In the hydrocarbon-impacted environments, there was a higher relative abundance of genes associated with iron acquisition and metabolism, photosynthesis, aromatic compound degradation, dormancy, motility, regulation and nitrogen metabolism, collectively contributing to 25.3% of the overall dissimilarity (Table 6.3). Previous

studies have shown that hydrocarbon-impacted environments were typified by an overall increase in genes related to iron acquisition and metabolism, dormancy and sporulation, motility, metabolism of aromatic compounds and cell signalling (Smith *et al*., unpublished data). Thus, results from this study further support the characterisation of hydrocarbon impacted sites by these functional genes.

6.5 Conclusion

Our data indicates that metagenomic signatures can be used to distinguish between contaminant types, with agricultural impact and hydrocarbon impact samples producing discrete functional signatures. In the agriculturally impacted environments, these signatures included metabolisms associated with cofactors, virulence, phages, fatty acids, protein metabolism, carbohydrates, amino acids and clustering based subsystems. In the hydrocarbon-impacted environment, the distinguishing metabolic signatures were genes associated with iron acquisition and metabolism, photosynthesis, aromatic compound degradation, dormancy, motility, regulation and nitrogen metabolism. Our data also indicated that the agricultural impact led to a more active community overall when compared to hydrocarbon impact. This study provides important insights into the different responses microbial communities have based on contaminant type, and suggest further investigation is needed given the wide range of chemicals that are currently affecting ecosystem health.

6.6 Acknowledgements

The authors gratefully acknowledge the funding provided by the Australian Research Council. R. J. Smith is the recipient of a Flinders University Research Scholarship (FURS).

Table 6.2 Results of CAP analysis for phylum-level taxonomy associated with impacted metagenomes.

	Factor	**m**	**Allocation Success % (ratio correct:misclassified)**					**δ^2**	***P*-value**
			Oil	Agricultural	Pristine	Wastewater	Total		
Taxonomy	Phylum	7	100 (4:4)	80 (4:5)	0 (0:1)	0 (0:2)	66.67	0.99	0.07
Metabolism	Level 1	2	75 (3:4)	100 (5:5)	0 (0:1)	50 (1:2)	75	**0.86**	**0.008**

Table 6.3 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon and agricultural impacted environments. Average dissimilarity between the two groups is 2.07%. Only metabolisms that were consistent (i.e. Diss/SD > 1.4) are shown here. The larger value in each case (i.e. the potential indicator of that condition) is shown in bold.

	Avg. Abundance			
Metabolic processes	**Hydrocarbon-Impacted**	**Agricultural-Impacted**	**Diss/SD**	**Cum %**
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.08	**0.18**	1.55	11.99
Virulence, Disease and Defence	0.08	**0.18**	1.55	23.97
Phages, Prophages, Transposable elements, Plasmids	0.08	**0.18**	1.55	35.96
Fatty Acids, Lipids, and Isoprenoids	0.08	**0.18**	1.55	47.94
Iron acquisition and metabolism	**0.84**	0.78	1.85	54.47
Photosynthesis	**0.69**	0.68	1.57	58.19
Metabolism of Aromatic Compounds	**0.87**	0.84	1.79	61.64
Dormancy and Sporulation	**0.71**	0.68	1.45	64.98
Motility and Chemotaxis	**0.83**	0.8	1.96	68.02
Protein Metabolism	0.93	**0.96**	3.5	70.94
Regulation and Cell Signalling	**0.85**	0.83	2.18	76.72
Carbohydrates	0.97	**0.99**	3.66	79.49
Nitrogen Metabolism	**0.84**	0.82	1.58	84.28
Amino Acids and Derivatives	0.96	**0.98**	2.22	86.22
Clustering-based subsystems	0.97	**0.99**	1.51	87.77

Cut-off percentage = 90%, Diss=dissimilarity; SD=Standard Deviation; Cum %=cumulative percentage of contribution to overall dissimilarity, Avg. Abundance values are reported for square-root transformed data

Table S6.1 Summary of publicly available metagenomes used in this study.

MG-RAST ID	Description/Reference
4453064.3	Unconfined aquifer (Smith *et al*., 2011)
4453083.3	Confined aquifer (Smith *et al*., 2011)
4440984.3	Coorong sediment 1 (Jeffries *et al*., 2011a)
4441020.3	Coorong sediment 2 (Jeffries *et al*., 2011a)
4441021.3	Coorong sediment 3 (Jeffries *et al*., 2011a)
4441022.3	Coorong sediment 4 (Jeffries *et al*., 2011a)
4453082.3	Hydrocarbon contaminated foreshore (Smith *et al*., unpublished data)
4453072.3	Hydrocarbon contaminated biopile (Smith *et al*., unpublished data)
4449126.3	Biopiles 2006 (Yergeau *et al*., 2012)
4450729.3	Biopile 2005 (Yergeau *et al*., 2012)
4455295.3	Wastewater 1 (Albertsen *et al*., 2012)
4463936.3	Wastewater 2 (Albertsen *et al*., 2012)

Table S6.2 Contribution of metabolic hierarchial system 1 to the dissimilarity of the hydrocarbon and agricultural impacted environments. Shows all metabolisms, including inconsistent ones (i.e. Diss/SD < 1.4). Average dissimilarity between the two groups is 2.07%. Bold values show either the condition with the higher average abundance (i.e. a potential indicator of that condition) or Diss/SD ratios that are consistent (i.e. > 1.4).

	Avg. Abundance			
Metabolic processes	**Hydrocarbon-Impacted**	**Agricultural-Impacted**	**Diss/SD**	**Cum %**
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.08	**0.18**	**1.55**	11.99
Virulence, Disease and Defence	0.08	**0.18**	**1.55**	23.97
Phages, Prophages, Transposable elements, Plasmids	0.08	**0.18**	**1.55**	35.96
Fatty Acids, Lipids and Isoprenoids	0.08	**0.18**	**1.55**	47.94
Iron acquisition and metabolism	**0.84**	0.78	**1.85**	54.47
Photosynthesis	**0.69**	0.68	**1.57**	58.19
Metabolism of Aromatic Compounds	**0.87**	0.84	**1.79**	61.64
Dormancy and Sporulation	**0.71**	0.68	**1.45**	64.98
Motility and Chemotaxis	**0.83**	0.8	**1.96**	68.02
Protein Metabolism	0.93	**0.96**	**3.5**	70.94
Potassium Metabolism	**0.79**	0.77	0.79	73.85
Regulation and Cell signalling	**0.85**	0.83	**2.18**	76.72
Carbohydrates	0.97	**0.99**	**3.66**	79.49
Secondary Metabolism	0.75	0.75	1.39	81.98
Nitrogen metabolism	**0.84**	0.82	**1.58**	84.28
Amino Acids and Derivatives	0.96	**0.98**	**2.22**	86.22
Clustering-based subsystems	0.97	**0.99**	**1.51**	87.77
Cell Division	0.84	0.84	0.73	89.27
Miscellaneous	0.94	**0.95**	**2.11**	90.65

Diss=dissimilarity; SD=Standard Deviation; Cum %=cumulative percentage of contribution to overall dissimilarity, Avg. Abundance values are reported for square-root transformed data

Chapter 7

Microbial response to anthropogenic disturbances: A general discussion

7.1 Overview

Environmental microbial communities are integral players in ecosystem functioning (Larsen *et al*., 2012; Lawrence *et al*., 2012). Following the introduction of xenobiotics, microbial communities are able to swiftly react to change, meaning they are highly resilient and excellent biological indicators (Steube *et al*., 2009). Despite their importance, microbial communities are often overlooked and consequently, remain poorly understood (Treseder *et al*., 2012). For that reason, the research presented in this thesis was stimulated by the need to gain an increased understanding of how environmental microbial communities respond to contaminants, to produce particular metagenomic signatures. The reoccurring theme throughout this thesis has been that major shifts in structure and functionality of the resident microbial communities were observed in metagenomic profiles following environmental change. This final chapter will discuss the major findings of the thesis and address the results from each of the experimental chapters within the context of the specific thesis aims outlined in Chapter 1.

7.1.1 Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems

The data presented in Chapter 2 addressed the first aim of the thesis by examining to what extent the composition and functionality of the resident microbial communities varied between a confined and surface-influenced unconfined aquifer ecosystem. This research was conducted in Ashbourne aquifer system which is characterised by two aquifer ecosystems with separate recharge processes that arise from distinct water sources (Banks *et al*., 2006; Smith *et al*., 2011; Roudnew *et al*., 2012). The unconfined aquifer lies below a dairy farming region and, therefore, receives agricultural input from the overlying environment. The confined aquifer however,

has been isolated from the surface for approximately 1500 years, providing a baseline for which to compare the unconfined aquifer to (Banks *et al*., 2006). A fundamental shift in taxa was observed with an overrepresentation of *Rhodospirillales*, *Rhodocyclales*, *Chlorobia* and *Circovirus* in the unconfined aquifer, while *Deltaproteobacteria* and *Clostridiales* were overrepresented in the confined aquifer (Fig. 2.2). A shift in metabolic processes was also observed, with a relative overrepresentation of genes associated with antibiotic resistance (β-lactamase genes), lactose and glucose utilization and DNA replication were observed in the unconfined aquifer, while genes associated with flagella production, phosphate metabolism and starch uptake pathways were all overrepresented in the confined aquifer (Fig. 2.3). These differences were likely driven by the extent of exposure to contaminants and nutrient input between the two groundwater systems. However, when the groundwater metagenomes, predominantly bacterial, were compared to metagenomes from a variety of environments, including ocean, freshwater, animal gut and sediment, the unconfined and confined aquifer were taxonomically and metabolically more similar to each other than to any other environment (Fig. 2.4 and 2.5). This suggests that the groundwater ecosystems had provided specific niches for the evolution of unique microbial communities.

7.1.2 Confined aquifers as viral reservoirs

In Chapter 3, we addressed the third aim by constructing a viral community profile of the viral sequences obtained in the unconfined and confined aquifer ecosystems, to further investigate the signature seen in the previous chapter. We found that despite geographical proximity, the viral community inhabiting the confined aquifer did not resemble that of the unconfined aquifer, and was instead most similar to the viral sequences in the metagenomes from a reclaimed water sample in Florida (Fig. 3.1)

(Rosario *et al*., 2009b; Smith *et al*., 2011; Roudnew *et al*., 2012). This result contradicted the previous chapter, whereby the patterns in bacterial taxonomy observed in the confined and unconfined aquifer were more similar to each other than to any other environment (Fig. 2.4 and 2.5). The similarity between the confined aquifer and reclaimed water source could suggest similar selective pressures, such a similar pore size, are driving community composition, leading to a similarity in the overall viral metagenomic signatures.

The taxa contributing to the similarity between the confined and reclaimed water viruses was further investigated, and it was found that the similarity was driven by a high relative occurrence of the ssDNA viral groups *Circoviridae*, *Geminiviridae*, *Inoviridae* and *Microviridae* (Fig. 3.2 and 3.3). *Circoviridae*, *Geminiviridae*, *Inoviridae*, *Microviridae* and *Nanoviridae* are all small viruses, with diameters of 7-30 nm (Storey *et al*., 1989; Gibbs and Weiller, 1999; Gutierrez *et al*., 2004). Therefore the dominance of these viruses is consistent with reports that small viruses have the greatest potential for transport through aquifers (Yates, 2000). Furthermore, *Circoviridae*, *Geminiviridae* and *Nanoviridae* all contain plant or vertebrate pathogens (Gibbs and Weiller, 1999; Gutierrez *et al*., 2004), with *Circoviridae* known to have a broad host range (Victoria *et al*., 2009; Delwarta and Li, 2012) indicating this viral group could be a potential health risk to humans. The identification of small ssDNA viruses in 1500 year-old groundwater suggests once viruses have been introduced, they can remain stable for long periods of time and thus, influence the viral metagenomic signature of groundwater ecosystems

7.1.3 Effect of hydrocarbon impacts on the structure and functionality of marine foreshore microbial communities: A metagenomic analysis

From the deep to the shallow, interstitial pore water communities experience similar matrices, but different types and concentrations of environmental impacts. Thus, Chapter 4 addressed the second aim of the thesis by assessing another common environmental pollutant, hydrocarbon contamination, and the effect it had on the structure and function of the microbial communities residing in historically impacted marine beach pore water. This research was conducted on hydrocarbon contaminated material from a former oil refinery site in Australia. When we compared our hydrocarbon impacted sample to two non-impacted samples, a shift in taxa was seen, with an overrepresentation of *Pseudomonadales*, *Actinomycetales*, *Rhizobiales*, *Alteromonadales*, *Oceanospirillales* and *Burkholderiales* in the hydrocarbon impacted sample (Fig. 4.2), all of which have previously been associated with impacted sites (Marcial Gomes *et al*., 2008). In addition to taxonomy, an overrepresentation of metabolic processes including aromatic compound metabolism, nitrogen metabolism and stress response were observed in the hydrocarbon impacted sample (Fig. 4.3). More specifically however, the increased relative abundance of *Oceanospirillales*, as well as a relative increase in nutrient metabolism and hydrocarbon degrading genes, suggests that the microbial potential to degrade hydrocarbon is being enhanced by coastal/seawater interactions.

To determine how the historical contamination event affected the overall structure and function of the inhabiting microbial communities, our hydrocarbon impacted foreshore metagenome was compared to metagenomes from 9 other marine habitats. Rank abundance plots showed the hydrocarbon impacted foreshore community had mid-range diversity indicative of a stable and functionally redundant community that

has adapted to stress (Table 4.2). We suggest this pattern is driven by the constant input of nutrients and water from tidal and wave action, as well as the low level contact with contaminants in the seawater, which have kept the relevant degradation genes selected for and induced.

7.1.4 Determining the metabolic footprints of hydrocarbon degradation using multivariate analysis

In Chapter 5 we conducted a multivariate analysis to characterise the metabolic footprints associated with hydrocarbon-impacted and non-impacted sediments. The hydrocarbon impacted foreshore metagenome discussed in Chapter 4 was used in conjunction with 3 other hydrocarbon impacted datasets to represent hydrocarbon impacted-environments, while 5 datasets were used for non-impacted environments. Unconstrained Multi-dimensional scaling (MDS) and constrained canonical analysis of principle coordinates (CAP) showed a clear distinction between the two groups (Fig. 5.1 and 5.2), with a high relative abundance of genes associated with cofactors, virulence, phages and fatty acids were present in the non-impacted sediments, collectively accounting for 45.7% of the overall dissimilarity (Table 5.2). Conversely, a high relative abundance of genes associated with iron acquisition and metabolism, dormancy and sporulation, motility, metabolism of aromatic compounds and cell signalling were observed in the hydrocarbon-impacted sites, together accounting for 22.3% of the overall dissimilarity (Table 5.2). Taken together, these results suggest the majority of the separation between the two groups was explained by a reduction in non-essential metabolisms in the hydrocarbon-impacted sediments. Furthermore, this reduction in non-essential metabolisms was coupled with a subsequent increase in pathways essential to the utilization of carbon and to survival.

7.1.5 Towards elucidating the metagenomic signature for impacted environments

Following on from the data obtained in Chapter 5, we sought to generate an overall metagenomic signature for impacted environments using CAP and similarity percentage analysis (SIMPER) in Chapter 6. Three common forms of environmental pollution were used, hydrocarbon impacted, including samples from chapter 4, agricultural impacted, including the groundwater samples from chapter 2, and wastewater. These groups were used to generate metagenomic signatures for the potential use as biological indicators. Significant differences between the relative abundance of metabolic processes in the impacted environments were shown, however, only the hydrocarbon and agricultural impacted environments could be correctly and consistently distinguished suggesting the sample size for wastewater was too low for comparison (Table 6.2). The main distinguishing metabolic processes associated with agricultural impacted environments were genes associated with cofactors, virulence, phages and fatty acids, while the main distinguishing genes associated with hydrocarbon impacted sites were iron acquisition and metabolism, photosynthesis, aromatic compound degradation, dormancy and motility (Table 6.3). As seen in Chapter 2, these results suggest markedly different community responses can be observed, making it possible to generate signatures based on contaminant type.

Combined, Chapters 5 and 6 addressed the fourth aim of this thesis by assessing our *a priori* hypothesis that community structure shifts in response to introduced contaminants. We were able to identify distinct metabolic processes based on

contaminant type, thus providing novel insight into the relative influence of anthropogenic modification on ecosystem functioning.

7.2 Thesis Synthesis: Demonstration of microbial indicators for impacted environments

It has been proposed that metagenomic analysis yields the most quantitative and accurate view of the microbial world (von Mering *et al*., 2007; Biddle *et al*., 2008), allowing for the assessment and exploitation of microbial communities on an ecosystem level (Simon and Daniel, 2009). Although this technology has vastly increased our knowledge of microbes in environmental systems, the complex relationship between community composition and ecosystem functioning is still being elucidated (Zengler and Palsson, 2012). Recent studies have demonstrated that metagenomes derived from similar environments have similar metagenomic signatures (Dinsdale *et al*., 2008b; Gianoulis *et al*., 2009; Willner *et al*., 2009; Jeffries *et al*., 2011a), however the characterisation of community composition based on contaminant type is scarcely understood. This thesis aimed to generate metagenomic signatures for two common forms of pollution worldwide, agricultural and hydrocarbon, thereby increasing our understanding of microbial community responses to contaminant type.

Previous anthropogenic modification studies have shown that microbial communities respond positively to nutrient and chemical pollutants by increasing productivity; however the specifics involved in the alteration of community functionality had not been explored in depth (Nogales *et al*., 2011). Results from this thesis demonstrated that agricultural modification led to an increase in genes associated with cofactors, virulence, phages, fatty acids, protein metabolism, carbohydrates, amino acids and

clustering based subsystems. Thus, the overall metagenomic signature associated with agricultural impact was defined by a more active community, likely driven by an increase in nutrient availability. Alternatively, hydrocarbon impacted microbial communities were shown to be expending the majority of their energy scavenging key nutrients needed for the bioremediation on hydrocarbons, at the expense of other, more complex pathways and growth, indicative of a less active community. Overall, this thesis demonstrated that microbial communities inhabiting impacted environments exhibited markedly different community responses based on contaminant type.

Additionally, this thesis showed that the microbial community response to anthropogenic modification was evident across multiple levels of taxonomic and metabolic resolution. Previous studies have supported this trend in that anthropogenic disturbances have led to major shifts in microbial dynamics that become evident across multiple levels (Hemme *et al*., 2010; Jeffries *et al*., 2011a). However, the majority of screening studies tend to focus on finer scale resolution (Joergensen and Emmerling, 2006). This thesis, however, has demonstrated the ability to screen at both coarse and finer levels of taxonomic and metabolic resolution, leading to a more robust set of metagenomic signatures. Furthermore, while taxonomic shifts are important in the assessment of discrete contamination events, the metabolic processes form the overall metagenomic signature for the comparison of impacted environments.

This thesis provides a novel insight into how environmental change, in the form of introduced contaminants, affects the microbial consortia. This study highlights the complexity and flexibility of microbial communities inhabiting stressed

environments, by showing how pollution shift the taxonomy and metabolism of microbial communities. This increases our understanding of the role these organisms play in ecosystem functioning.

Although high-throughput sequencing platforms have revolutionized the field of microbial ecology, the major limiting factor for information density and accuracy are computational power and error profiles associated with the different platforms. For example, the error rate associated with the 454 GS FLX Titanium sequencer is in the range of 10^{-3} – 10^{-4}, which is lower than the other new, high-throughput sequencing platforms such as Illumina and SOLiD (Kircher and Kelso, 2010). As sequencing platforms and computational power increase however, our ability to characterize complete communities, beyond that of the most dominant species, will continue to improve. Increased sensitivity within sequencing technologies will also reduce the yield of DNA required, thus reducing and eliminating the need for biased amplification steps. Advances in molecular technologies and computational power coupled with cell enumeration protocols and environmental metadata, would produce a thorough understanding of how current changes in environmental conditions are effecting our planet.

References

Abbaszadegan M, Lechevallier M, Gerba C (2003) Occurrence of viruses in U.S. groundwaters. J Am Water Works Assoc 95: 107-120.

Ager D, Evans S, Li H, Lilley AK, van der Gast CJ (2010) Anthropogenic disturbance affects the structure of bacterial communities. Environ Microbiol 12: 670-678.

Albertsen M, Benedict Skov Hansen L, Marc Saunders A, Halkjær Nielsen P, Lehmann Nielsen K (2012) A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. ISME J 6: 1094-1106.

Al-Zabet T (2002) Evaluation of aquifer vulnerability to contamination potential using the DRASTIC method. Environ Geol 43: 203-208.

Anderson MJ, Robinson J (2001) Permutation tests for linear models. Aust NZ J Stat 43: 75-88.

Anderson MJ, Willis TJ (2003) Canonical analysis of principle coordinates: A useful method of constrained ordination for ecology. Ecology 84: 511-525.

Anderson MJ, Gorley RN, Clarke KR (2008) PERMANOVA+ for PRIMER: Guide to software and statistical methods. PRIMER-E, Plymouth UK.

Anderson T (2003) Microbial eco-physiological indicators to asses soil quality. Agric Ecosyst Environ 98: 285-293.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F (2006) The marine viromes of four oceanic regions. PLoS Biol 4: 2121-2131.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schnieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Michael Miller R, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comp Biol 5: e1000593.

APHA (2005) Standard methods for the examination of water and wastewater, 21st edn. Washington, DC, USA: American Public Health Association.

Asner GP, Elmore AJ, Olander LP, Martin RE, Harris AT (2004) Grazing systems, ecosystem responses and global change. Annu Rev Environ Resour 29: 261-299.

Atlas RM (1995) Petroleum biodegradation and oil spill bioremediation. Mar Pollut Bull 31: 178-182.

Atlas RM, and Bartha R (1972) Degradation and mineralization of petroleum in sea water: limitation by nitrogen and phoshorous. Biotechnol Bioeng 14: 309-318.

Atlas RM, and Hazen TC (2011) Oil biodegradation and bioremediation: A tale of the two worst spills in U.S. history. Environ Sci Technol 45: 6709-6715.

Avidano L, Gamalero E, Cossa GP, Carraro E (2005) Characterization of soil health in an Italian polluted site by using microorganisms as bioindicators. Appl Soil Ecol 30: 21-33.

Bailey JK, Wooley SC, Lindroth RL, Whitham TG (2006) Importance of species interactions to community heritability: a genetic basis to trophic-level interactions. Ecol Lett 9: 78-85.

Baker KL, Langenheder S, Nicol GW, Ricketts D, Killham K, Campbell CD, Prosser JI (2009) Environmental and spatial characterisation of bacterial community composition in soil to inform sampling strategies. Soil Biol Biochem 41: 2292-2298.

Balser TC, McMahon KD, Bart D, Bronson D, Coyle DR, Craig N, Flores-Mangual ML, Forshay K, Jones SE, Kent AE, Shade AL (2006) Bridging the gap between micro - and macro-scale perspectives on the role of microbial communities in global change ecology. Plant Soil 289: 59-70.

Banks EW, Wilson T, Green G, Love A (2006) Groundwater recharge investigations in the Eastern Mount Lofty Ranges, South Australia, Report DWLBC 2007/20,

Government of South Australia, through Department of Water, Land and Biodiversity Conservation, Adelaide.

Barnosky AD, Hadly EA, Bascompte J, Berlow EL, Brown JH, Fortelius M, Getz, WM, Harte J, Hastings A, Marquet PA, Martinez NA, Mooers A, Roopnarine P, Vermeij G, Williams JW, Gillespie R, Kitzes J, Marshell C, Matzke N, Mindell DP, Revilla E, Smith AB (2012) Approaching a state shift in Earth's biosphere. Nature 486: 52-58.

Bass Becking LGM (1934) Geobiologie of inleiding tot de milieukunde. The Hague, the Netherlands: WP Van Stockum & Zoon (in Dutch).

Bastias BA, Huang ZQ, Blumfield T, Xu Z, Cairney JWG (2006) Influence of repeated prescribed burning on the soil fungal community in an eastern Australian wet sclerophyll forest. Soil Biol Biochem 38: 3492-3501.

Beller HR, Grbic-Galic D, Reinhard M (1992) Microbial degradation of toluene under sulfate-reducing conditions and the influence of iron on the process. Appl Environ Microbiol 58: 786-793.

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE (2007) Metagenomic characterization of chesapeake bay virioplankton. Appl Environ Microbiol 73: 7629-7641.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc B (Methodological) 57: 289-300.

Bennett PC, Rogers JR, Choi WJ (2001) Silicates, silicate weathering, and microbial ecology. Geomicrobiol J 18: 3-19.

Berga M, Székely AJ, Langenheder S (2012) Effects of disturbance intensity and frequency on bacterial community composition and function. PLos One 7: e36959.

Berghash SR, Davidson JN, Armstrong JC, Dunny GM (1983) Effects of antibiotic treatment of nonlactating dairy cows on antibiotic resistance patterns of bovine mastitis pathogens. Antimicrob Agents Ch 24: 771-776.

Biddle JF, Fitz-Gibbon S, Schuster SC, Brenchley JE, House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. Proc Nat Acad Sci USA 105: 10583-10588.

Binga EK, Lasken RS, Neufeld JD (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. ISME J 2: 233-241.

Bond NR, Lake PS, Arthington AH (2008) The impacts of drought on freshwater ecosystems: an Australian perspective. Hydrobiologia 600: 3-16.

Bonjoch X, Ballesté E, Blanch AR (2004) Multiplex PCR with 16S rRNA gene-targeted primers of *Bifidobacterium* spp. to identify sources of fecal pollution. Applied and Environ Microbiol 70: 3171-3175.

Borchardt MA, Bradbury KR, Gotkowitz MB, Cherry JA, Parker BL (2007) Human enteric viruses in groundwater from a confined bedrock aquifer. Environ Sci Technol 41: 6606-6612.

Brulc JM, Antonopoulos DA, Berg Miller ME, Wilson MK, Yannarell AC, Dinsdale EA, Edwards RE, Frank ED, Emerson JB, Wacklin P, Coutinho PM, Henrissat B, Nelson KE, White BA (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc Nat Acad Sci USA 106: 1948-1953.

Brune A, Frenzel P, Cypionka H (2000) Life at the oxic-anoxic interface: microbial activities and adaptations. FEMS Microbiol Rev 24: 691-710.

Brussaard CPD (2004) Optimization of procedures for counting viruses by flow cytometry. Appl Environ Microbiol 70: 1506-1513.

Buckley DH, Schmidt TM (2003) Diversity and dynamics of microbial communities in soils from agro-ecosystems. Environ Microbiol 5: 441-452.

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2011) Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci USA 108: 14288-14293.

Buyer JS, Drinkwater LE (1997) Comparison of substrate utilization assay and fatty acid analysis of soil microbial communities. J Microbiol Methods 30: 3-11.

Carpenter SR, Stanley EH, Vander Zanden MJ (2011) State of the world's freshwater ecosystems: physical, chemical, and biological changes. Annu Rev Environ Resour 36: 75-99.

Chang Y, Peacock AD, Long PE, Stephen JR, McKinley JP, Macnaughton SJ, Anwar Hussain AKM, Saxton AM, White DC (2001) Diversity and characterization of sulfate-reducing bacteria in groundwater at a uranium mill tailings site. Appl Environ Microbiol 67: 3149-3160.

Chakraborty R, Wu CH, Hazen TC (2012) Systems biology approach to bioremediation. Curr Opin Biotechnol 23: 1-8.

Chapin III FS, Walker BH, Hobbs RJ, Hooper DU, Lawton JH, Sala OE, Tilman D (1997) Biotic control over the functioning of ecosystems. Science 277: 500-504.

Chapelle FH, Lovley DR (1990) Rates of microbial metabolism in deep coastal plain aquifers. Appl Environ Microbiol 56: 1865-1874.

Chikere CB, Okpokwasili GC, Chikere BO (2011) Monitoring of microbial hydrocarbon remediation in the soil. 3 Biotech 1: 117-138.

Cho J, Kim S (2000) Increase in bacterial community diversity in subsurface aquifers receiving livestock wastewater input. Appl Environ Microbiol 66: 956-965.

Clarke KR (1993) Nonparametric multivariate analysis of changes in community structure. Aust J Ecol 18: 117-143.

Clarke KR, Gorley R (2006) PRIMER v6: User Manula/Tutorial. Plymouth, UK: PRIMER-E.

Clarke KR, Warwick RM (2001) Change in Marine Communities: An Approach to Statistical Analysis and Interpretation, 2nd Edition. PRIMER-E Ltd.

Cookson WR, Osman M, Marschner P, Abaye DA, Clark I, Murphy DV Stockdale EA, Watson CA (2007) Controls on soil nitrogen cycling and microbial community composition across land use and incubation temperature. Soil Biol Biochem 39: 744-756.

Crutsinger GM, Collins MD, Fordyce JA, Gompert Z, Nice CC, Sanders NJ (2006) Plant Genotypic Diversity Predicts Community Structure and Governs an Ecosystem Process. Science 313: 966-968.

Dandie CE, Weber J, Aleer S, Adetutu EM, Ball AS, Juhasz AL (2010) Assessment of five bioaccessibility assays for predicting the efficacy of petroleum hydrocarbon biodegradation in aged contaminated soils. Chemosphere 81: 1061-1068.

Danielopol DL, Pospisil P, Rouch R (2000) Biodiversity in groundwater: A large-scale view. Trends Ecol Evol 15: 223-224.

Danielopol DL, Griebler C, Gunatilaka A, Notenboom J (2003) Present state and future prospects for groundwater ecosystems. Enviro Conserv 30: 104-130.

de Wit R, Bouvier T (2006) 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? Environ Microbiol 8: 755-758.

Delille D, Delille B (2000) Field observations on the variability of crude oil impact on indigenous hydrocarbon-degrading bacteria from sub-Antarctic intertidal sediments. Mar Environ Res 49: 403-417.

DeLong EF (2009) The microbial ocean from genomes to biomes. Nature 459: 200-206.

DeLong EF, Pace NR (2001) Environmental diversity of bacteria and archaea. Syst Biol 50: 470-478.

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, Martinez A, Sullivan MB, Edwards R, Rodriguez Brito B, Chisholm SW, Karl DM (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.

Delwarta E, Li L (2012) Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. Virus Res 164: 114-121.

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Vega Thurber R, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature Letters 452: 340-345.

Diels L (2005) Immobilised cell strategies for the treatment of soil and groundwaters. Applications of cell immobilisation biotechnology 8B: 495-506.

Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L, Hatay M, Hall D, Brown E, Haynes M, Krause L, Sala E, Sandin SA, Vega Thurber R, Willis BL, Azam F, Knowlton N, Rohwer R (2008a) Microbial ecology of four coral atolls in the northern line islands. PLoS One 3: 1-17.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, Furlan M, Desnues C, Haynes M, Li L, McDaniel L, Moran MA, Nelson KE, Nilsson C, Olson

R, Paul J, Britto BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Vega Thurber R, Wegley L, White BA, Rohwer F (2008b) Functional metagenomic profiling of nine biomes. Nature 452: 629-633.

Donlan RM (2002) Biofilms: microbial life on surfaces. Emerg Infect Dis 8: 881-890.

Eaton TT, Anderson MP, Bradbury KR (2007) Fracture control of ground water flow and water chemistry in a rock aquitard. Groundwater 45: 601-615.

Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504-510.

Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander Jr EC, Rohwer F (2006) Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics 7: 1-13.

Ekendahl S, O'Neill AH, Thomsson E, Pedersen K (2003) Characterisation of yeasts isolated from deep igneous rock aquifers of the fennoscandian shield. Microbial Ecol 46: 416-428.

Fernandez AS, Hashsham SA, Dollhopf SL, Raskin L, Glagoleva O, Dazzo FB, Hickey RF, Criddle CS, Tiedje JM (2000) Flexible community structure correlates with stable community function in methanogenic bioreactor communities perturbed by glucose. Appl Environ Microbiol 66: 4058-4067.

Fernández-Luqueño F, Valenzuela-Encinas C, Marsch R, Martínez-Suárez C, Vázquez-Núñez E, Dendooven L (2011) Microbial communities to mitigate contamination of PAHs in soil—possibilities and challenges: a review. Environ Sci Pollut R 18: 12-30.

Fisher WD (1958) On grouping for maximum homogeneity. J Amer Statist Ass 53 789-798.

Foley JA, DeFries R, Asner GP, Barford C, Bonan G, Carpenter SR, Stuart Chapin F, Coe MT, Daily GC, Gibbs HK, Helkowski JH, Holloway T, Howard EA, Kucharik CJ, Monfreda C, Patz JA, Prentice IC, Ramankutty N, Snyder PK (2005) Global consequences of land use. Science 309: 570-574.

Follows MJ, Dutkiewicz S, Grant S, Chisholm SW (2007) Emergent biogeography of microbial communities in a model ocean. Science 315: 1843-1846.

Ford TE (2000) Response of marine microbial communities to anthropogenic stress. J Aquat Ecosyst Stress Recovery 7: 75-89.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP (2009) The bacterial species challenge: Making sense of genetic and ecological diversity. Science 323: 741-746.

Fry NK, Fredrickson JK, Fishbain S, Wagner M, Stahl DA (1997) Population structure of microbial communities associated with two deep, anaerobic, alkaline aquifers. Appl Environ Microbiol 63: 1498-1504.

Fuhrman JA (2009) Microbial community structure and its functional implications. Nature 459: 193-199.

Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S (2006) Annually reoccurring bacterial communities are predictable from ocean conditions. Proc Nat Acad Sci USA 103: 13104-13109.

Gallego JLR, Loredo J, Llamas JF, Vázquez F, Sánchez J (2001) Bioremediation of diesel-contaminated soils: Evaluation of potential in situ techniques by study of bacterial degradation. Biodegradation 12: 325-335.

Ge Y, Zhang J, Zhang L, Yang M, He J (2008) Long-term fertilization regimes affect bacterial community structure and diversity of an agricultural soil in northern China. J Soils Sediments 8: 43-50.

Gelsomino A, Badalucco L, Ambrosoli R, Crecchio C, Puglisi E, Meli SM (2006) Changes in chemical and biological soil properties as induced by anthropogenic disturbance: A case study of an agricultural soil under recurrent flooding by wastewaters. Soil Biol Biochem 38: 2069-2080.

Gasol JM, Del Giorgio PA (2000) Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. Scientia Marina 64: 197-224.

Ghosh S, LaPara TM (2007) The effects of subtherapeutic antibiotic use in farm animals on the proliferation and persistence of antibiotic resistance among soil bacteria. ISME J 1: 191-203.

Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. Proc Natl Acad Sci USA 106: 1374-1379.

Gillooly JF, Allen AP, Savage VM, West GB (2004) Toward a metabolic theory of ecology. Ecology 85: 1771-1789.

Gianneechini RE, Concha C, Franklin A (2002) Antimicrobial susceptibility of udder pathogens isolated from dairy herds in the west littoral region of uruguay. Acta Vet Scan 43: 31-41.

Gibbs MJ, Weiller GF (1999) Evidence that a plant virus switched hosts to infect a vertebrate and then recombined with a vertebrate-infecting virus. Proc Natl Acad Sci USA 96: 8022-8027.

Gibert J, Deharveng L (2002) Subterranean ecosystems: A truncated functional biodiversity. Bioscience 52: 473-481.

Gonzalez A, King A, Robeson MS, Song S, Shade A, Metcalf JL, Knight R (2012) Characterizing microbial communities through space and time. Curr Opin Biotechnol 23: 431-436.

Goodall DW (1964) A probabilistic similarity index. Nature 203: 1098.

Goodall DW (1966) A new similarity index based on probability. Biometrics 22: 882-907.

Griebler C, Lueders T (2009) Microbial biodiversity in groundwater ecosystems. Freshwat Biol 54: 649-677.

Griffiths BS, Bonkowski M, Roy J, Ritz K (2001) Functional stability, substrate utilisation and biological indicators of soils following environmental impacts. Appl Soil Ecol 16: 49-61.

Griffiths BS, Philippot L (2012) Insights into the resistance and resilience of the soil microbial community. FEMS Microbiol Ecol DOI:10.1111/j.1574-6976.2012.00343.x.

Guerrero R, Montesinos E, Pedrós-Alió C, Esteve I, Mas J, van Gemerden H, Hofman PAG, Bakker JF (2002) Phototrophic sulfur bacteria in two Spanish lakes: Vertical distribution and limiting factors. Limnol Oceanogr 30: 919-931.

Gutierrez C, Ramirez-Parra E, Mar Castellano M, Sanz-Burgos AP, Luque A, Missich R (2004) Geminivirus DNA replication and cell cycle interactions. Vet Microbiol 98: 111-119.

Haberl H, Erb KH, Krausmann F, Gaube V, Bondeau A, Plutzar C, Gingrich S, Lucht W, Fischer-Kowalski M (2007) Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. Proc Nat Acad Sci USA 104: 12942-12947.

Hall-Stoodley L, Costerton WJ, Stoodley P (2004) Bacterial biofilms: From the natural environment to infectious diseases. Nat Rev Microbiol 2: 95-108.

Hamamura N, Olson SH, Ward DM, Inskeep WP (2006) Microbial population dynamics associated with crude-oil biodegradation in diverse soils. Appl Environ Microbiol 72: 6316-6324.

Hamblin KW, Christiansen EH (2004) Earth's dynamic systems, 10th Ed. In. New Jersey, USA: Pearson Education Inc.

Hara A, Syutsubo K, Harayama S (2003) Alcanivorax which prevails in oil-contaminated seawater exhibits broad substrate specificity for alkane degradation. Environ Microbiol 5: 746-753.

Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortnew JL, Stringfellow WT, Bill M, Conrad ME, Tom

LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D'haeseller P, Holman HN, Osman S, Lu Z, Van Nostrand JD, Deng Y, Zhou J, Maso OU (2010) Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science 330: 204-208.

Head IM, Jones DM, Röling WFM (2006) Marine microorganisms make a meal of oil. Nat Rev Microbiol 4: 173-182.

Hemme CL, Deng Y, Gentry TJ, Fields MW, Wu L, Barua S, Barry K, Tringe SG, Watson DB, He Z, Hazen TC, Tiedje JM, Rubin EM, Zhou J (2010) Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. ISME J 4: 660-672.

Henriques IDS, Aga DS, Mendes P, Connor SKO, Love NG (2007) Metabolic footprinting: A new approach to identify physiological changes in complex microbial communities upon exposure to toxic chemicals. Environ Sci Technol 41: 3945-3951.

Hernesmaa A, Björklöf K, Kiikkilä O, Fritze H, Haahtela K, Romantschuk M (2005) Structure and function of microbial communities in the rhizosphere of Scots pine after tree-felling. Soil Biol Biochem 37: 777-785.

Hesselsoe M, Füreder S, Schloter M, Bodrossy L, Iversen N, Roslev P, Halkjær Nielsen P, Wagner M, Loy A (2009) Isotope array analysis of *Rhodocyclales* uncovers functional redundancy and versatility in an activated sludge. ISME J 3: 1349-1364.

Holmes DE, Risso C, Smith JA, Lovley DR (2011) Anaerobic oxidation of benzene by the hyperthermophilic archaeon *Ferroglobus placidus*. Appl Environ Microbiol 77: 5926-5933.

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377-386.

Jeffries TC, Seymour JR, Gilbert JA, Dinsdale EA, Newton K, Leterme SSC, Roudnew B, Smith RJ, Seuront L, Mitchell JG (2011a) Substrate type determines metagenomic profiles from diverse chemical habitats. PLoS One 6:e25173. DOI:25110.21371/journal.pone.0025173.

Jeffries TC, Seymour JR, Newton K, Smith RJ, Seuront L, Mitchell JG (2011b) Increases in the abundance of microbial genes encoding halotolerance and photosynthesis along a sediment salinity gradient. Biogeosciences 8: 7551-7574.

Joergensen RG, Emmerling C (2006) Methods for evaluating human impact on soil microorganisms based on their activity, biomass, and diversity in agricultural soils. J Plant Nutr Soil Sci 169: 295-309.

Johnsen AR, Karlson U (2005) PAH degradation capacity of soil microbial communities—does it depend on PAH exposure? Microbial Ecol 50: 488-495.

Karl DM (2002) Nutrient dynamics in the deep blue sea. Trends Microbiol 10: 410-418.

Kaye JP, McCulley RL, Burke IC (2005) Carbon fluxes, nitrogen cycling, and soil microbial communities in adjacent urban, native and agricultural ecosystems. Global Change Biol 11: 575-587.

Ke L, Yu KSH, Wong YS, Tam NFY (2005) Spatial and vertical distribution of polycyclic aromatic hydrocarbons in mangrove sediments. Sci Total Environ 340: 177-187.

Keller M, Hettich R (2009) Environmental proteomics: a paradigm shift in characterizing microbial activities at the molecular level. Microbiol Mol Biol R 73: 62-70.

Kemper N (2008) Veterinary antibiotics in the aquatic and terrestrial environment. Ecol Indic 8: 1-13.

Kennedy J, Flemer B, Jackson SA, Lejon DPH, Morrissey JP, O'Gara F, Dobson ADW (2010) Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. Mar Drugs 8: 608-628.

Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. BioEssays 32: 524-536.

Kleikemper J, Schroth MH, Sigler WV, Schmucki M, Bernasconi SM, Zeyer J (2002) Activity and diversity of sulfate-reducing bacteria in a petroleum hydrocarbon-contaminated aquifer. Appl Environ Microbiol 68: 1516-1523.

Klijn N, Weerkamp AH, Nizo WMDV (1995) Detection and characterization of lactose-utilizing *Lactococcus* spp. in natural ecosystems. Appl Environ Microbiol 61: 788-792.

Kostka JE, Prakash O, Overholt WA, Green SJ, Freyer G, Canion A, Delgardio J, Norton N, Hazen TC, Huettel M (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: 7962-7974.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P (2010) Wrinkles in the rare biosphere: pyrosequencing errors lead to artificial inflation of diversity estimates. Environ Microbiol 12: 118-123.

Kurokawa K, Itoh T, Kuwahara Y, Ashima K, Toh H, Toyoda A, Takami H, Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y, Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M (2007) Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res 14: 169-181.

Langworthy DE, Stapleton RD, Sayler GS, Findlay RH (1998) Genotypic and phenotypic responses of a riverine microbial community to polycyclic aromatic hydrocarbon contamination. Appl Environ Microbiol 64: 3422-3428.

Larsen PE, Field D, Gilbert JA (2012) Predicting bacterial community assemblages using an artificial neural network approach. Nature Methods 9: 621-625.

Lawrence D, Fiegna F, Behrends V, Bundy JG, Phillimore AB, Bell T, Barraclough TG (2012) Species interactions alter evolutionary responses to a novel environment. PLoS Biol 10: e1001330.

Leahy JG, Colwell RR (1990) Microbial degradation of hydrocarbons in the environment. Microbiol Rev 54: 305-315.

Lear G, Lewis GD (2009) Impact of catchment land use on bacterial communities within stream biofilms. Ecol Indic 9: 848-855.

Leclerc H, Moreau A (2002) Microbiological safety of natural mineral water. FEMS Microbiol Rev 26: 207-222.

Lee E, Sacks WJ, Chase TN, Foley JA (2011) Simulated impacts of irrigation on the atmospheric circulation over Asia. J Geophys Res 116: DO8114.

Lefeuvre P, Lett JM, Varsani A, Martin DP (2009) Widely conserved recombination patterns among single-stranded DNA viruses. J Virol 83: 2697-2707.

Liang Y, Van Nostrand JD, Wang J, Zhang X, Zhou J, Li G (2009) Microarray-based functional gene analysis of soil microbial communities during ozonation and biodegradation of crude oil. Chemosphere 75: 193-199.

Liang Y, Van Nostrand JD, Deng Y, He Z, Wu L, Zhang X, Li G, Zhou J (2011) Functional gene diversity of soil microbial communities from five oil-contaminated fields in China. ISME J 5: 403-413.

Liebana E, Batchelor M, Hopkins KL, Clifton-Hadley FA, Teale CJ, Foster A, Barker L, Threlfall EJ, Davies RH (2006) Longitudinal farm study of extended-spectrum Beta-Lactamase-mediated resistance. J Clin Microbiol 44: 1630-1634.

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009) High diversity of the viral community from an Antarctic lake. Science 326: 858-861.

Loreau M, Naeem S, Inchausti P, Bengtsson J, Grime JP, Hector A, Hooper DU, Huston MA, Raffaelli D, Schmid B, Tilman D, Wardle DA (2001) Biodiversity and ecosystem functioning: Current knowledge and future challenges. Science 294: 804-808.

Loy A, Schulz C, Lücker S, Schöpfer-Wendels A, Stoecker K, Baranyi C, Lehner A, Wagner M (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "*Rhodocyclales*". Appl Environ Microbiol 71: 1373-1386.

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proc Natl Acad Sci USA 104: 11436-11440.

Macnaughton SJ, Stephen JR, Venosa AD, Davis GA, Chang Y, White DC (1999) Microbial population changes during bioremediation of an experimental oil spill. Appl Environ Microbiol 65: 3566-3574.

Mäder P, Fliebβach A, Dubois D, Gunst L, Fried P, Niggli U (2002) Soil fertility and biodiversity in organic farming. Science 296: 1694-1697.

Madigan MT, Martinko JM, Dunlap PV, Clark DP (2003) Brock biology of microorganisms, 10th edn. New Jersey: Pearson Education Inc.

Madison LL, Huisman GW (1999) Metabolic engineering of Poly(3-Hydroxyalkanoates): From DNA to plastic. Microbiol Mol Biol R 63: 21-53.

Maila MP, Cloete TE (2005) The use of biological activities to monitor the removal of fuel contaminants—perspective for monitoring hydrocarbon contamination: a review. Int Biodeterior 55: 1-8.

Malone T, Davidson M, DiGiacomo P, Gonçalves E, Knap T, Muelbert J, Parslow J, Sweijd N, Yanagai T, Yap H (2010) Climate change, sustainable development and coastal ocean information needs. Procedia Environ Sci 1: 324-341.

Männistö MK, Tiirola MA, Salkinoja-Salonen MS, Kulomaa MS, Puhakka JA (1999) Diversity of chlorophenol-degrading bacteria isolated from contaminated boreal groundwater. Arch Microbiol 171: 189-197.

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, Morales R, Allen AE, Armbrust EV (2012) Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability Proc Nat Acad Sci USA 109: E317–E325.

Marcial Gomes NC, Borges LR, Paranhos R, Pinto FN, Mendonça-Hagler LCS, Smalla K (2008) Exploring the diversityof bacterial communities in sediments of urban mangrove forests. FEMS Microbiol Ecol 66: 96-109.

Margesin R, Schinner F (2001) Bioremediation (natural attenuation and biostimulation) of diesel-oil-contaminated soil in an alpine glacier skiing area. Appl Environ Microbiol 67: 3127-3133.

Marie D, Partensky F, Jacquet S, Vaulot D (1997) Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR green I. Appl Environ Microbiol 63: 186-193.

Marie D, Brussaard CPD, Thyrhaug R, Bratbak G, Vaulot D (1999) Enumeration of marine viruses in culture and natural samples by flow cytometry. Appl Environ Microbiol 65: 45-52.

Maron P, Mougel C, Ranjard L (2011) Soil microbial diversity: Methodological strategy, spatial overview and functional interest. C R Biol 334: 403-411.

Martinez-Garcia M, Swan BK, Poulton NJ, Lluesma Gomez M, Masland D, Sieracki ME, Stepanauskas R (2012) High-throughput single-cell sequencing identifies photoheterotrophs and chemoautotrophs in freshwater bacterioplankton. ISME J 6: 113 - 123.

Martins dos Santos VAP, Sabirova J, Timmis KN, Yakimov MM, Golyshin PN (2010) Alcanivorax borkumensis. Handbook of Hydrocarbon and Lipid Microbiology Part 14: 1265-1288.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, Horner-Devine MC, Kane M, Krumins JA, Kuske CR, Morin PJ, Naeem S, Øvreås L, Reysenbach A, Smith VH, Staley JT (2006) Microbial biogeography: putting microorganisms on the map. Nature Reviews 4: 102-112.

Marzorati M, Wittebolle L, Boon N, Daffonchio D, Verstraete W (2008) How to get more out of molecular fingerprints: practical tools for microbial ecology. Environ Microbiol 10: 1571-1581.

Mathews CK (2006) DNA precursor metabolism and genomic stability. FASEB J 20: 1300-1314.

Meyer F, Paarmann D, D'Souza M, Olsem R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 1-8.

Mitchell JG (2004) Rank-size analysis and vertical phytoplankton distribution patterns. Handbook of Scaling Methods in Aquatic Ecology: Measurement, Analysis, Simulation P G Strutton and L Seuront eds CRC Press, Boca Raton: 257-278.

Mitra S, Gilbert JA, Field D, Huson DH (2010) Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. ISME J 4: 1236-1242.

Mpelasoka F, Hennessy K, Jones R, Bates B (2008) Comparison of suitable drought indices for climate change impacts assessment over Australia towards resource management. Int J Climatol 28: 1283-1292.

Naeem S, Li S (1997) Biodiversity enhances ecosystem reliability. Nature 390: 507-509.

Nayar GPS, Hamel AL, Lin L, Sachvie C, Grudeski C, Spearman G (1999) Evidence for circovirus in cattle with respiratory disease and from aborted bovine fetuses. Can Vet J 40: 277-278.

Nazir J, Haumacher R, Ike A, Stumpf P, Böhm R, Marschang RE (2010) Long-Term study on tenacity of avian influenza viruses in water (distilled water, normal saline, and surface water) at different temperatures. Avian Dis 54: 720-724.

Neufeld JD, Chen Y, Dumont MG, Murrell JC (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. Environ Microbiol 10: 1526-1535.

Newcombe RG (1998) Improved confidence intervals for the difference between binomial proportions based on paired data. Stat Med 17: 2635-2650.

Nikolopoulou M, Kalogerakis N (2008) Biostimulation strategies for fresh and chronically polluted marine environments with petroleum hydrocarbons. J Chem Technol and Biotechnol 84: 802-807.

Nogales B, Lanfranconi MP, Piña-Villalonga JM, Bosch R (2011) Anthropogenic perturbations inmarine microbial communities. FEMS Microbiol Rev 35: 275-298.

Nolan BT, Ruddy BC, Hitt KJ, Helsel DR (1997) Risk of nitrate in groundwaters of the United Statess - A national perspective. Environ Sci Technol 31: 2229-2236.

Ortega-Calvo JJ, Marchenko AI, Vorobyov AV, Borovick RV (2003) Chemotaxis in polycyclic aromatic hydrocarbon-degrading bacteria isolated from coal-tar- and oil-polluted rhizospheres. FEMS Microbiol Ecol 44: 373-381.

Ouellette GD, Buckley PE, O'Connell KP (2010) Environmental influences on the relative stability of baculoviruses and *Vaccinia* virus: A review. Emerging Endemic Pathogens 00: 125-149.

Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang H, Cohoon M, de Crécy-Lagard V, Diaz N, Disz N, Edwards R, Fonstein M, Frank ED, Gerdes S, Glass EM, Goesmann A, Hanson A, Iwata-Reuyl D, Jensen R, Jamshidi N, Krause L, Kubal M, Larsen N, Linke B, McHardy AC, Meyer F, Neuweger H, Olsen G, Olson R, Osterman A, Portnoy V, Pusch GD, Rodionov DA, Rückert C, Steiner J, Stevens R, Thiele I, Vassieva O, Ye Y, Zagnitko O, Vonstein V (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691-5702.

Palenik B, Ren Q, Tai V, Paulsen IT (2009) Coastal synechococcus metagenome reveals major roles for horizontal gene transfer and plasmids in population diversity. Environ Microbiol 11: 349-359.

Paredes-Sabja D, Setlow P, Sarker MR (2011) Germination of spores of *Bacillales* and *Clostridiales* species: mechanisms and proteins involved. Trends Microbiol 19: 85-94.

Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics 26: 715-721.

Parnell JJ, Crowl TA, Weimer BC, Pfrender ME (2009) Biodiversity in microbial communities: system scale patterns and mechanisms. Mol Ecol 18: 1455-1462.

Pedersen K (1993) The deep subterranean biosphere. Earth-Sci Rev 34: 243-260.

Pedersen K (2000) Exploration of deep intraterrestrial microbial life: current perspectives. FEMS Microbiol Lett 185: 9-16.

Pelz O, Tesar M, Wittich R, Moore ERB, Timmis KN, Abraham W (1999) Towards elucidation of microbial community metabolic pathways: unravelling the network of carbon sharing in a pollutant-degrading bacterial consortium by immunocapture and isotopic ratio mass spectrometry. Environ Microbiol 1: 167-174.

Peng R, Xiong A, Xue Y, Fu X, Gao F, Zhao W, Tian Y, Yao Q (2008) Microbial biodegradation of polyaromatic hydrocarbons. FEMS Microbiol Rev 32: 927-955.

Qu A, Brulc JM, Wilson MK, Law BF, Theoret JR, Joens LA, Konkel ME, Angly F, Dinsdale EA, Edwards RA, Nelson KE, White BA (2008) Comparative metagenomics reveals host specific metavirulomes and horizontal gene transfer elements in the chicken cecum microbiome. PLoS One 3: e2945.

Ramette A, Tiedje JM (2007) Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microbial Ecol 53: 197-207.

Ramšak A, Peterka M, Tajima K, Martin JC, Wood J, Johnston MEA, Aminov RI, Flint HJ, Avguštin G (2000) Unravelling the genetic diversity of ruminal bacteria belonging to the CFB phylum. FEMS Microbiol Ecol 33: 69-79.

Rittman BE, Krajmalnik-Brown R, Halden RU (2008) Pre-genomic, genomic and post-genomic study of microbial communities involved in bioenergy. Nature Reviews Microbiol 6: 604-612.

Rogers JR, Bennett PC (2004) Mineral stimulation of subsurface microorganisms: release of limiting nutrients from silicates. Chem Geol 203: 91-108.

Röling WFM, Milner MG, Martin Jones D, Lee K, Daniel F, Swannell RJP, Head IM (2002) Robust hydrocarbon degradation and dynamics of bacterial communities during nutrient-enhanced oil spill bioremediation. Appl Environ Microbiol 68: 5537-5548.

Röling WFM, Milner MG, Martin Jones D, Fratepietro F, Swannell RPJ, Daniel F, Head IM (2004) Bacterial community dynamics and hydrocarbon degradation during a field-scale evaluation of bioremediation on a mudflat beach contaminated with buried oil. Appl Environ Microbiol 70: 2603-2613.

Röling WFM, Ferrer M, Golyshin PN (2010) Systems approaches to microbial communities and their functioning. Curr Opin Biotechnol 21: 532-538.

Ron EZ, Rosenberg E (2002) Biosurfactants and oil bioremediation. Curr Opin Biotechnol 13: 249-252.

Rosenberg E, Legmann R, Kushmaro A, Taube R, Adler E, Ron EZ (1992) Petroleum bioremediation - a multiphase problem. Biodegradation 3: 337-350.

Rosenberg E (2006) Hydrocarbon-oxidizing bacteria. Prokaryotes 2: 564-577.

Rosario K, Duffy S, Breitbart M (2009a) Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90: 2418-2424.

Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M (2009b) Metagenomic analysis of viruses in reclaimed water. Environ Microbiol 11 2806-2820.

Roudnew B, Seymour JR, Jeffries TC, Lavery TJ, Smith RJ, Mitchell JG (2012) Bacterial and virus-like particle abundances in purged and unpurged groundwater depth profiles. Groundwater Monit R DOI: 10.1111/j1745–6592.2012.01393.x.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JF, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Stausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Biology 5: 398-431.

Sabirova J (2010) Polyhydroxyalkanoates produced by hydrocarbon-degrading bacteria. Handbook of Hydrocarbon and Lipid Microbiology 27: 2981-2994.

Sanapareddy N, Hamp TJ, Gonzalez LC, Hilger HA, Fodor AA, Clinton SM (2008) Molecular diversity of a north carolina wastewater treatment plant as revealed by pyrosequencing. Appl Environ Microbiol 75: 1688-1696.

Santos HF, Carmo FL, Paes JES, Rosado AS, Peixoto RS (2011) Bioremediation of mangroves impacted by petroleum. Water Air Soil Poll 216: 329-350.

Sawant AA, Sordillo LM, Jayarao BM (2005) A survey on antibiotic usage in dairy herds in Pennsylvania. J Dairy Sci 88: 2991-2999.

Scheuerman PR, Farrah SR, Bitton G (1987) Reduction of microbial indicators and viruses in a cypress strand. Water Sci Technol 19: 539-546.

Schimel J, Balser TC, Wallenstein M (2007) Microbial stress-response physiology and its implications for ecosystem function. Ecology 88: 1386-1394.

Schneiker S, Martins dos Santos VAP, Bartels D, Bekel T, Brecht M, Buhrmester J, Chernikova TN, Denaro R, Ferrer M, Gertler C, Goesmann A, Golyshina OV, Kaminski F, Khachane AN, Lang S, Linke B, McHardy AC, Meyer F, Nechitaylo T, Pühler A, Regenhardt D, Rupp O, Sabirova JS, Selbitschka W, Yakimov MM, Timmis KN, Vorhölter F, Weidner S, Kaiser O, Golyshin PN (2006) Genome sequence of the ubiquitous hydrocarbon-degrading marine bacterium Alcanivorax borkumensis. Nature Biotechnol 24: 997-1004.

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol 74: 4164- 174.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A Community Resource for Metagenomics. PLoS Biol 5: 394-397.

Seymour JR, Doblin MA, Jeffries TC, Brown MV, Newton K, Ralph PJ, Baird M, Mitchell JG (2012) Contrasting microbial assemblages in adjacent water-masses associated with the East Australian Current. Environ Microbiol Rep DOI: 10.1111/j.1758-2229.2012.00362.x.

Sheppard PJ, Adetutu EM, Makadia TH, Ball AS, (2011) Microbial community and ecotoxicity analysis of bioremediated, weathered hydrocarbon-contaminated soil. Soil Res 49: 261-269

Sherr EB, Sherr BF (1991) Planktonic microbes: Tiny cells at the base of the ocean's food webs. Trends Ecol Evol 6: 50-54.

Short JW, Irvine GV, Mann DH, Maselko JM, Pella JJ, Lindeberg MR, Payne JR, Driskell WB, Rice SD (2007) Slightly weathered *Exxon Valdez* oil persists in gulf of Alaska beach sediments after 16 years. Environ Sci Technol 41: 1245-1250.

Siciliano SD, Germida JJ, Banks K, Greer CW (2003) Changes in microbial community composition and function during a polyaromatic hydrocarbon phytoremediation field trial. Appl Environ Microbiol 69: 483-489.

Sim Y, Chrysikopoulos CV (2000) Virus transport in unsaturated porous media. Water Resour Res 36: 173-179.

Simon C, Daniel R (2009) Achievements and new knowledge unraveled by metagenomic approaches. Appl Microbiol Biotechnol 85: 265-276.

Singh AK, Sherry A, Gray ND, Jones MD, Röling WFM, Head IM (2011) Dynamics of *Alcanivorax* spp. in oil-contaminated intertidal beach sediments undergoing bioremediation. Appl Microbiol Mol Biol DOI 10.1007/978-90-481-9252-6_24: 199-209.

Singleton I (1994) Microbial metabolism of xenobiotics: Fundamental and applied research. J Chem Technol Biotechnol 59: 9-23.

Sjöstedt J, Koch-Schmidt P, Pontarp M, Canbäck B, Tunlid A, Lundberg P, Hagström Å, Riemann L (2012) Recruitment of members from the rare biosphere of marine bacterioplankton communities after an environmental disturbance. Appl Environ Microbiol 78: 1361–1369.

Smith RJ, Jeffries TC, Roudnew B, Fitch AJ, Seymour JR, Delpin MW, Newton K, Brown MH, Mitchell JG (2011) Metagenomic comparison of microbial communities

inhabiting confined and unconfined aquifer ecosystems. Environ Microbiol 14: 240-253.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrita JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored ‘‘rare biosphere’’. Proc Natl Acad Sci USA 103: 12115–12120.

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 28: 1409-1438.

Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY, Chow CT, Sachdeva R, Jones AC, Schwalbach MS, Rose JM, Hewson I, Patel A, Sun F, Caron DA, Fuhrman JA (2011) Marine bacterial, archaeal and protistan association networks reveal ecological linkages. ISME J 5: 1414-1425.

Stegen JC, Lin X, Konopka AE, Fredrickson JK (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. ISME J DOI:10.1038/ismej.2012.22.

Steube C, Richter S, Griebler C (2009) First attempts towards an integrative concept for the ecological assessment of groundwater ecosystems. Hydrogeol J 17: 23-35.

Storey CC, Lusher M, Richmond SJ, Bacon J (1989) Further characterization of a bacteriophage recovered from an avian strain of *Chlamydia psittaci*. J Gen Virol 70: 1321-1327.

Storey JD, Tibshirani R (2003) SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In Parmigiani,G, Garrett ES, Irizarry RA, Zeger SL (eds), The Analysis of Gene Expression Data: Methods and Software. Springer, New York.

Suenaga H, Ohnuki T, Miyazaki K (2007) Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. Environ Microbiol 9: 2289-2297.

Sun B, Dong Z, Zhang X, Li Y, Cao H, Cui Z (2011) Rice to vegetables: Short- versus long-term impact of land-use change on the indigenous soil microbial community. Microbial Ecol 62: 474-485.

Swannell RPJ, Croft BC, Grant AL, Lees K (1995) Evaluation of bioremediation agents in beach microcosms. Spill Sci Technol B 2: 151-159.

Swindoll MC, Aelion MC, Pfaender FK (1988) Influence of inorganic and organic nutrients on aerobic biodegradation and on the adaptation response of subsurface microbial communities. Appl Environ Microbiol 54: 212-217.

Teixeira de Mattos MJ, Neijssel OM (1997) Bioenergetic consequences of microbial adaptation to low-nutrient environments. J Biotechnol 59: 117-126.

Toze S (1999) PCR and the detection of microbial pathogens in water and wastewater. Water Res 33: 3545-3556.

Treseder KK, Balser TC, Bradford MA, Brodie EL, Dubinsky EA, Eviner VT, Hofmockel KS, Lennon JT, Levine UY, MacGregor BJ, Pett-Ridge J, Waldrop MP (2012) Integrating microbial ecology into ecosystem models: challenges and priorities. Biogeochemistry 109: 7-18.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communites. Science 308: 554-557.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI (2009) A core gut microbiome in obese and lean twins. Nature Letters 457: 480-485.

Urich T, Lanzén A, Qi J, Huson DH, Schleper C, Schuster SC (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3: 1-13

van der Heijden MGA, Bardgett RD, van Straalen NM (2008) The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett 11: 296-310.

van der Zaan BM, Saia FT, Stams AJM, Plugge CM, De Vos WM, Smidt H, Langenhoff AAM, Gerritse J (2012) Anaerobic benzene degradation under denitrifying conditions: *Peptococcaceae* as dominant benzene degraders and evidence for a syntrophic process. Environ Microbiol DOI:10.1111/j.1462-2920.2012.02697.x.

Venkata Mohan S, Kisa T, Ohkuma T, Kanaly RA, Shimizu Y (2006) Bioremediation technologies for treatment of PAH-contaminated soil and strategies to enhance process efficiency. Rev Environ Sci Biotechnol 5: 347-374.

Venosa AD, Suidan MT, Wrenn BA, Strohmeier KL, Haines JR, Loye Eberhart B, King D, Holder E (1996) Bioremediation of an experimental oil spill on the shoreline of Delaware Bay. Environ Sci Technol 30: 1764-1775.

Venosa AD, Zhu X (2003) Biodegradation of crude oil contaminating marine shorelines and freshwater wetlands. Spill Sci Technol Bull 8: 163-178.

Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, Naeem A, Zaidi S, Delwart E (2009) Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. J Virol 83: 4642-4651.

Vinas M, Sabaté J, José Espuny M, Solanas AM (2005) Bacterial community dynamics and polycyclic aromatic hydrocarbon degradation during bioremediation of heavily creosote-contaminated soil. Appl Environ Microbiol 71: 7008-7018.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315: 1126-1130.

Vreeland RH, Piselli AF, Jr, McDonnough S, Meyers SS (1998) Distribution and diversity of halophilic bacteria in a subsurface salt formation. Extremophiles 2**:** 321–331.

Walworth J, Pond A, Snape I, Rayner J, Ferguson S, Harvey P (2007) Nitrogen requirements for maximizing petroleum bioremediation in a sub-Antarctic soil. Cold Reg Sci Technol 48: 84-91.

Warnecke F, Luginbühl P, Ivanova N, Ghassemian M, Richardson TH, Stege JT, Cayouette M, McHardy AC, Djordjevic G, Aboushadi N, Sorek R, Tringe SG, Podar M, Martin HG, Kunin V, Dalevi D, Madejska J, Kirton E, Platt D, Szeto E, Salamov A, Barry K, Mikhailova N, Kyrpides NC, Matson EG, Ottensen EA, Zhang X, Hernández M, Murillo C, Acosta LG, Rigoutsos I, Tamayo G, Green BD, Chang C, Rubin EM, Mathur EJ, Robertson DE, Hugenholtz, Leadbetter JR (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature 450: 560-569.

Watanabe K (2001) Microorganisms relevant to bioremediation. Curr Opin Biotechnol 12: 237-241.

White DC, Flemming CA, Leung KT, Macnaughton SJ (1998) In situ microbial ecology for quantitative appraisal, monitoring, and risk assessment of pollution

remediation in soils, the subsurface, the rhizosphere and in biofilms. J Microbiol Methods 32: 93-105.

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. Proc Natl Acad Sci USA 95: 6578-6583.

Williams BK, Titus K (1988) Assessment of sampling stability in ecological applications of discriminant analysis. Ecology 69: 1275-1285.

Willner D, Thurber RV, Rohwer F (2009) Metagenomic signatures of 86 microbial and viral metagenomes. Environ Microbiol 11: 1752-1766.

Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: Read length matters. Appl Environ Microbiol 74: 1453–1463.

Wooley JC, Ye Y (2010) Metagenomics: Facts and artifacts, and computational challenges. J Comput Sci Technol 25: 71-81.

Woyke T, Xie G, Copeland A, González JM, Han C, Kiss H, Saw JH, Senin P, Yang C, Chatterji S, Cheng J, Eisen JA, Sieracki, Stepanauskas R (2009) Assembling the marine metagenome, one cell at a time. PLoS One 4: e5299.

Wu Y, Luo Y, Zou D, Ni J, Liu W, Teng Y, Li Z (2008) Bioremediation of polycyclic aromatic hydrocarbons contaminated soil with *Monilinia* sp.: degradation and microbial community analysis. Biodegradation 19: 247-257.

Xing M, Zhang X, Huang H (2012) Application of metagenomic techniques in mining enzymes from microbial communities for biofuel synthesis. Biotechnol Adv 30: 920-929.

Xu R, Obbard JP, Tay ETC (2003) Optimization of slow-release fertilizer dosage for bioremediation of oil-contaminated beach sediment in a tropical environment. World J Microbiol Biotechnol 19: 719-725.

Yakimov MM, Timmis KN, Golyshin PN (2007) Obligate oil-degrading marine bacteria. Curr Opin Biotechnol 18: 257-266.

Yates MV, Gerba CP, Kelley LM (1985) Virus persistence in groundwater. Appl Environ Microbiol 49: 778-781.

Yates MV (2000) Measurement of virus and indicator survival and transport in subsurfaces. AWWA Res Found.

Yergeau E, Arbour M, Brousseau R, Juck D, Lawrence JR, Masson L, Whyte LG, Greer CW (2009) Microarray and real-time PCR analyses of the responses of high-arctic soil bacteria to hydrocarbon pollution and bioremediation treatments. Appl Environ Microbiol 75: 6258-6267.

Yergeau E, Sanschagrin S, Beaumier D, Greer CW (2012) Metagenomic analysis of the bioremediation of diesel-contaminated Canadian high arctic soils. PLoS One 7: e30058.

Yilmaz S, Allgaier, M, Hugenholtz P (2010) Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature Methods 7: 943-944.

Zengler K, Palsson BO (2012) A road map for the development of community systems (CoSy) biology. Nature Reviews Microbiol 10: 366-372.

Zhang X, Zhang T, Fang HHP (2009) Antibiotic resistance genes in water environment. Appl Microbiol Biotechnol 82: 397-414.

Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Mark Welch DB, Martiny JBH, Sogin M, Boetius A, Ramette A (2011) Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS One 9: e24570.

Appendix 1

Published Manuscripts Arising From and Related to this Thesis