Mol Biotechnol (2013) 54:900–912
DOI 10.1007/s12033-012-9639-z

RESEARCH

Analysis of Hydrocarbon-Contaminated Groundwater Metagenomes as Revealed by High-Throughput Sequencing

Nathlee S. Abbai · Balakrishna Pillay

Published online: 11 January 2013

Abstract The tendency for chlorinated aliphatics and aromatic hydrocarbons to accumulate in environments such as groundwater and sediments poses a serious environmental threat. In this study, the metabolic capacity of hydrocarbon (aromatics and chlorinated aliphatics)-contaminated groundwater in the KwaZulu-Natal province of South Africa has been elucidated for the first time by analysis of pyrosequencing data. The taxonomic data revealed that the metagenomes were dominated by the phylum *Proteobacteria* (mainly *Betaproteobacteria*). In addition, *Flavobacteriales*, *Sphingobacteria*, *Burkholderiales*, and *Rhodocyclales* were the predominant orders present in the individual metagenomes. These orders included microorganisms (*Flavobacteria*, *Dechloromonas aromatica* RCB, and *Azoarcus*) involved in the degradation of aromatic compounds and various other hydrocarbons that were present in the groundwater. Although the metabolic reconstruction of the metagenome represented composite cell networks, the information obtained was sufficient to address questions regarding the metabolic potential of the microbial communities and to correlate the data to the contamination profile of the groundwater. Genes involved in the degradation of benzene and benzoate, heavy metal-resistance mechanisms appeared to provide a survival strategy used by the microbial communities. Analysis of the pyrosequencing-derived data revealed that the metagenomes represent complex microbial communities that have adapted to the geochemical conditions of the groundwater as evidenced by the presence of key enzymes/genes conferring resistance to specific contaminants. Thus, pyrosequencing analysis of the metagenomes provided insights into the microbial activities in hydrocarbon-contaminated habitats.

Keywords Hydrocarbon-contaminated metagenomes · Taxonomic assignments · Metabolic profiles

Introduction

Groundwater environments can hold the largest pool of microorganisms in the biosphere with estimates of bacterial abundances reaching 3.8–6.0 $\times$ 10^{30} cells [37]. However, the complete characterization of microbial communities in natural systems remains a challenge due to their high diversity and the uncultured status of approximately 99 % of the microbial population [14]. However, the emergence of metagenomic techniques has provided researchers with the tools to resolve the full extent of the uncultured microbial diversity as well as allowed access to the biochemical pathways within these uncultured microorganisms [19].

High-throughput pyrosequencing technology, which does not require cloning or PCR amplification, has been recently applied in the metagenomic characterization of environmental microbial communities [36]. This approach has increased the number and scope of metagenomic sequencing projects [16]. Sequencing of microbial communities from several environments, including acid-mine drainage, marine water, sediments, and soil have provided insights regarding novel gene discovery, metabolism, community structure, function, and evolution [14] and thus opens a new window to the hidden world of microbial communities. The sites investigated in this

Electronic supplementary material The online version of this article (doi:10.1007/s12033-012-9639-z) contains supplementary material, which is available to authorized users.

N. S. Abbai · B. Pillay (✉)
Discipline of Microbiology, Faculty of Science and Agriculture, University of KwaZulu-Natal (Westville Campus), Private Bag X54001, Durban 4000, Republic of South Africa
e-mail: pillayb1@ukzn.ac.za

study were within an industrial plant which has been in operation for the past 60 years. During this time, an array of anthropogenic pollutants (chlorinated aliphatics and aromatics) have been deposited at the plant with seepage into the subsurface environments which could potentially contaminate the surrounding potable water source posing serious problems on human health. The duration, cost, and uncertain success of pumping out of these contaminants are daunting obstacles both to the protection of groundwater and to the recycling and re-development of industrially contaminated land [31]. However, microorganisms have been shown to evolve an extensive range of enzymes and pathways that are able to degrade a wide array of xenobiotics [22].

In this study, we investigated the taxonomic and metabolic profiles of the resident microbial communities from two hydrocarbon-contaminated groundwater samples to gain insights into the adaptation of the microbial communities to severe environmental contamination. In addition, mechanisms for the detoxification of environmental contaminants from the sequence data were predicted.

Materials and Methods

Site Sampled

Groundwater samples were collected from two boreholes, designated Borehole 1 and Borehole 2 within a contaminated site from KwaZulu-Natal, South Africa, in April 2009. This site has been in operation since 1907, during which time a diverse range of chemicals including chlorine, herbicides, pesticides, paints, and explosives have been manufactured. In addition, an assessment of the studied site conducted by Palmer et al. [30] showed that the waste deposited at the site is comprised mainly of chlorinated hydrocarbons, a number of inorganics, predominantly BTEX (benzene, toluene, ethylbenzene, and xylene). A total volume of 5 L was collected from a depth of 15 m and transferred to sterile Schott bottles. Upon collection, the samples were transported to the laboratory at room temperature. The pH and temperature of the water measured on site was shown to be 5.96 and 25 °C, respectively.

Analysis of the Contaminated Water Samples

The groundwater was analyzed to determine the level of volatile organic compounds (VOC) as well as the major cations present in the groundwater. The VOC analysis (Table 1) was carried out in duplicate by purge and trap GC–MS at the Council of Scientific and Industrial Research (CSIR) in Modderfontein, South Africa. The presence of major cations was performed by CLEAN STREAM Scientific Services (Pty) Ltd., Pretoria, South Africa, and the data are represented in Table 2. A paired *t* test was used to compare the level of contamination (VOC) and chemical properties (cations) of the two Boreholes.

Isolation of Total DNA

A total volume of 5 L for each sample was filtered using a 47-mm Advantec water filtration system (Toyo Roshi Kaisha, Ltd.) fitted with a 0.22-μm nylon filter (Satorius). Filters were then washed overnight in PBS at 4 °C. Following washing, the filters were vortexed to dislodge the sample matrix from the filter and centrifuged to pellet the sample material. The filters were discarded and the DNA isolation was performed on the resultant pellet. DNA was isolated using the UltraClean Soil DNA Kit (MoBio Laboratories, Inc.). Numerous DNA isolation kits were tested; however, the best performance was recorded with the above-mentioned kit. The resulting DNA preparations were quantified and checked for purity using the Nanodrop 1000 Spectrophotometer (Thermo Scientific). Owing to the low microbial biomass in groundwater systems, we used multiple displacement amplification before pyrosequencing. This method has been used widely to amplify DNA before sequencing [37]. In this study, GenomiPhi was used on both water samples analyzed here, therefore any bias in the process is applied to both samples.

Pyrosequencing and Analysis of GS FLX Titanium Data

DNA was sequenced (70 × 75 mm PicoTitrePlate) on a Roche GS FLX titanium pyrosequencer. Pyrosequencing library construction and sequencing was performed as described by Margulies et al. [25]. The raw reads obtained were assembled into contigs using the CLC Genomics de novo assembly tool default settings (Aarhus, Denmark). Raw 454 reads were directly exported to the database. Following sequencing, simple contig sequences were created using all the information which was in the read sequences. This was the actual de novo part of the process. These simple contig sequences do not contain any information about which reads the contigs are built from. Second, all the reads were mapped using the simple contig sequence as reference. This was done to show, e.g., coverage levels along the contigs and enabling more downstream analysis like SNP detection and creating mapping reports. The output of the assembly was not a graph, but a list of contig sequences. When all the previous optimization and scaffolding steps had been performed, a contig sequence was produced for every non-ambiguous path in the graph. If the path was not fully resolved, *N*s were inserted as an estimation of the distance between two nodes.

Table 1 VOC analysis of contaminated water samples

Sample ID	Borehole 1 (μg/L)	Borehole 2 (μg/L)	Sample ID	Borehole 1 (μg/L)	Borehole 2 (μg/L)
Dichlorodifluoromethane	<10	<10	Dibromochloromethane	<10	<10
Vinyl chloride	790	2800	1,2-Dibromoethane	<10	<10
Trichlorofluoromethane	<10	<10	Chlorobenzene	83	<10
1,1-Dichloroethene	1300	<10	1,1,1,2-Tetrachloroethane	<10	<10
Dichloromethane	<40	<40	Ethylbenzene	<10	<10
trans-1,2-Dichloroethene	40	24	*m,p*-Xylene	<20	<20
1,1-Dichloroethane	210	<10	*o*-Xylene	<10	<10
cis-1,2-Dichloroethene	2100	1300	Styrene	<10	<10
2,2-Dichloropropane	<10	<10	Bromoform	<10	<10
Bromochloromethane	<10	<10	Isopropylbenzene	<10	<10
Chloroform	120	<10	1,1,2,2-Tetrachloroethane	19	46
1,1,1-Trichloroethane	<10	<10	1,2,3-Trichloropropane	15	<10
1,1-Dichloropropene	<10	<10	Bromobenzene	<10	<10
Carbon tetrachloride	<10	<10	*n*-Propylbenzene	<10	<10
1,2-Dichloroethane	480	39	2-Chlorotoluene	<10	<10
Trichloroethene	390	260	1,3,5-Trimethylbenzene	<10	<10
1,1,1,2-Tetrachloroethane	<10	<10	4-Chlorotoluene	<10	<10
Benzene	25	<10	*tert*-Butylbenzene	<10	<10
1,2-Dichloropropane	<10	<10	1,2,4-Trimethylbenzene	<10	<10
Dibromomethane	<10	<10	*sec*-Butylbenzene	<10	<10
Bromodichloromethane	<10	<10	4-Isopropyltoluene	<10	<10
1,1,2-Trichloroethane	1900	<10	1 1,3-Dichlorobenzene	<10	<10
1,3-Dichloropropane	<10	<10	1,4-Dichlorobenzene	<10	<10
1,2-Dibromo-3-chloropropane	<10	<10	*n*-Butylbenzene	<10	<10
1 1,2,4-Trichlorobenzene	<10	<10	1,2-Dichlorobenzene	10	<10
Hexachlorobutadiene	12	<10	Naphthalene	<10	<10
			1,2,3-Trichlorobenzene	<10	<10

Table 2 Major cations present in the contaminated samples

Cation (mg/L)	Site 1	Site 2
Calcium (Ca)	80.278	473.515
Magnesium (Mg)	22.003	807.58
Sodium (Na)	359.35	1000.24
Potassium (K)	5.371	13.992
Aluminum (Al)	<0.006	<0.006
Iron (Fe)	0.028	0.035
Manganese (Mn)	0.012	5.402
Total chromium (Cr)	0.002	0.025
Copper (Cu)	0.028	0.31
Nickel (Ni)	0.02	0.348
Zinc (Zn)	0.009	0.082
Cobalt (Co)	<0.002	<0.002
Cadmium (Cd)	<0.001	<0.001
Lead (Pb)	0.03	0.29

The assembled data was annotated using the MetaGenome Rapid Annotation using Subsystem Technology (MG-RAST) [28]. In addition, the raw reads were imported into MG-RAST; however, the data output obtained were not different from the data obtained with the assembled reads. MG-RAST performs a normalization step for all uploaded data. This step involves the generation of unique internal IDs and the removal of exact duplicate sequences from 454 datasets (These sequences are an artifact of the sequencing technique and are not scientifically meaningful). The procedure included two steps which were applied, independently, to each metagenomic sample: transformation and standardization. After each sample had undergone transformation and standardization, the values for all considered samples were scaled from 0 (the minimum value of all considered samples) to 1 (the maximum value of all considered samples). This was a uniform scaling that does not affect the relative differences

of values within a single sample or between/among two or more samples. This procedure placed all values on a scale from 0 to 1, and was used to produce figures where the entire abundance range (for all samples under consideration) was expressed on a scale from 0 to 1. This eliminates negative abundance values presenting all abundance counts in a more intuitive scale. The sequences were then screened for potential protein encoding genes by means of a BLASTX search against the SEED comprehensive non-redundant database. All BLAST searches were performed using an expected value of 1×10^{-5} and a minimum alignment length of 50 bp. At this cutoff, 3 of the observed hits would be expected to occur at random [2].

Results and Discussion

Analysis of the Groundwater Samples

Data obtained for the VOC and major cation analysis of the samples are represented in Tables 1 and 2, respectively. According to the VOC analysis, low to moderate concentrations of aromatics, such as benzene, and chlorinated aliphatic hydrocarbons (CAHs), such as vinyl chloride (VC), *trans*-1,2-dichloroethene, *cis*-1,2-dichloroethene, 1,2-dichloroethane (DCA), trichloroethene (TCE) and tetrachloroethene (PCE), were detected at the individual sites (Table 1). However, there was no significant difference ($p = 0.377$) in the concentrations of the organic compounds present in the individual boreholes. We can only assume that the higher concentration of VC (2,800 μg/L) observed in Borehole 2 when compared to borehole 1 (790 μg/L) may be indicative of biodegradation of chlorinated hydrocarbons by the resident microbial community in that sample. This would also suggest that the microbial community harbors catabolic pathways for CAHs [9]. In addition, the low concentration observed for benzene (<10 μg/L) also suggested the presence of possible aromatic degradative pathways. However, the amplification of benzene-degradative genes by gene-specific PCR merits future research, which will contribute to more conclusive evidence regarding the presence of benzene catabolic genes. However, VOC analysis of Borehole 1 revealed higher concentrations of PCE, TCE, and *cis*-1,2-DCE, but a lower concentration of VC, suggesting that the resident microbial community is yet to acquire catabolic genes for the CAHs or that the genes may be present, but their expression requires an external source for stimulation (e.g., vitamins and nutrients). However, all these assumptions need to be validated with the sequence data. Although the cationic composition of the groundwater indicated that the calcium concentration was 5-fold, magnesium 36-fold, sodium 2-fold, and potassium 2-fold higher in Borehole 2 than in Borehole 1 (Table 2), statistical analysis revealed that there was no significance difference ($p = 0.09$) in the levels of major cations in both the samples. Therefore, the direct contribution of the cations in the sample that may act as biostimulants for degradation of the contaminants cannot be confirmed. However, by analyzing the taxonomic distribution and metabolic potential of the microorganisms present in the individual boreholes, a more holistic view of the processes taking place at the sites will be achieved.

Overview of the Metagenomic Sequencing

Pyrosequencing generated data was used to analyze the composition and metabolic potential of the microbial communities present in the individual metagenomes. The dataset for Borehole 1 (MG-RAST ID: 4450733.3) contained 4,573 sequences totaling 3,967,039 basepairs with an average length of 867 bps. Zero sequences (0.0 %) failed to pass the QC pipeline. From the dataset, 3,139 sequences (68.6 %) contain predicted proteins with known functions and 1,426 sequences (31.2 %) contain predicted proteins with unknown function. Likewise, the dataset for Borehole 2 (MG-RAST ID: 4450734.3) contained 8,091 sequences totaling 4,524,932 base pairs with an average length of 559 bps. Zero sequences (0.0 %) failed to pass the QC pipeline. Of the sequences that passed QC, 45 sequences (0.6 %) contain ribosomal RNA genes. Of the remainder, 4,694 sequences (58.0 %) contain predicted proteins with known functions and 3,297 sequences (40.7 %) contain predicted proteins with unknown function.

Phylogenetic Composition of the Bacterial Community

The phylogenetic diversity of the metagenome was assessed by the evaluation of similarities to conserved protein families and domains in the pyrosequencing data set using MG-RAST's underlying SEED database [28] and CARMA [21]. CARMA involves the characterization of species composition and the genetic potential of microbial samples using short reads. In contrast to the traditional 16S-rRNA approach for taxonomical classification, CARMA uses reads which encode for known proteins. By assigning the taxonomic origins to each read, a profile is constructed which characterizes the taxonomic composition of the corresponding community.

A shift in dominant taxa was seen between Borehole 1 and Borehole 2 with different communities inhabiting each environment. Based on the data obtained, bacterial sequences dominated the boreholes with the majority of the sequences belonging to the phylum *Bacteriodetes* and *Proteobacteria* repectively. In addition, the current study employed further phylogenetic analysis by amplifying the bacterial 16S rRNA variable regions, conducting denaturing gradient gel electrophoresis (DGGE) analysis and

sequencing of 15 of the most dominant DGGE bands (data not shown). However, it was observed that sequence data obtained from a 200-bp product is not adequate for providing an absolute phylogenetic affiliation; it is a mere estimation of the microbial population present in that community. The data obtained with CARMA [21] confirmed the sequence data obtained with the SEED analysis (Figs. S1, S2).

Comparisons of the sequences to MG-RASTs underlying SEED database revealed that Borehole 1 was dominated by the *Bacteriodetes* (Fig. 1a) to which most of the sequences belonged to *Flavobacteriales* and *Sphingobacteria* (Fig. 2a). The dominant phylum present in Borehole 2 was *Proteobacteria* (Fig. 1b) of which most of the proteobacterial sequences were assigned to the class *Betaproteobacteria*. In addition, the majority of the bacterial sequences that dominated Borehole 2 belonged to the orders *Burkholderiales* and *Rhodocyclales* (Fig. 2b).

A study conducted by Schmeisser et al. [34] revealed that a drinking water biofilm 16S rRNA clonal library contained a vast majority of clones belonging to the bacterial phylotype *Flavobacterium-Bacteroides* group, as evidenced in this study. In subsurface communities when nutrient levels are low, microbes attach themselves to sediment particles and biofilms. Therefore, microbes dominating groundwater systems are commonly found attached to surfaces rather than being in suspension [37]. Within, the current study's dataset sequences contributing to biofilm formation were identified such as exopolysaccharides (rhamnose-containing glycans). During the sampling process, it was possible that the biofilms were dislodged, thus being present in the sample which was processed.

However, of interest to the current study, was data reported by [13] who highlighted the hydrocarbon degradative properties of *Flavobacterium* isolates. In addition, previous studies reported on the bioremediation of polycyclic aromatic hydrocarbons using a bacterial consortium [18, 23] with *Sphingobacteria* shown to be one of the dominant microorganisms present in these environments. In this study, Borehole 1 was also shown to be a dominated by *Sphingobacteria*.

A study conducted by Aburto et al. [1], which investigated benzene biodegradation in hydrocarbon-contaminated groundwater, showed that the most abundant phylogenetic group of *Bacteria* in the most contaminated groundwater was the *Betaproteobacteria.* In this study, Borehole 2 was dominated by *Betaproteobacteria* and this was evidenced by an over-representation of *Burkholderiales* and *Rhodocyclales* (Fig. 2b). *Burkholderiales* were shown to dominate a stressed groundwater microbial community which had been exposed to high concentrations of heavy metals, nitric acid, and organic solvents for about 50 years [14].

Similarly, *Rhodocyclales* were shown to be over-represented in an unconfined groundwater system which was shown to contain an abundance of resistance genes (antibiotic resistance). In addition, *Rhodocyclales* are well known for their ability to degrade aromatic compounds [37]. Previous studies have shown that microorganisms such as *Azoarcus* sp. EbN1 and *Dechloromonas aromatica* RCB, which belong to this order, have been implicated in the degradation of hydrocarbons present in contaminated sites, which is characteristic of the groundwater samples in this study [3, 32, 35]. In addition, the files containing data obtained by protein recruitment plots showed the greatest similarity to *Dechloromonas aromatica* RCB and *Azoarcus* sp. EbN1 (Additional Files 1 and 2 in Supplementary Material). The recruitment plot tool was used to provide a selected sequenced microbial genome as a scaffold to map metagenome-derived sequences to only sequences which were annotated from a metagenome used as the queries. The initial view provided a ranked list of microbial genomes that contained the most number of matched sequences from the metagenome. This gave an indication of the relative representations in terms of genomic content found within the metagenome. These results generated from the recruitment plots showed the link between the contamination profile of the groundwater samples and dominant microorganisms present. Previous studies [3, 32, 35] have reported that *Dechloromonas aromatica* and *Azoarcus* possess enzymes which make up pathways involved in the degradation of hydrocarbons [11], *Azoarcus evansii* has also been shown to utilize benzoate, 3-hydroxybenzoate, and gentisate aerobically as sole sources of carbon and energy sources [3]. In addition, studies conducted by [42] have also confirmed the ability of *Azoarcus* to utilize benzoate as the sole carbon and energy source. Therefore, it is possible that a similar scenario is taking place in the groundwater communities. However, there is still a need for further investigation regarding this assumption.

Strains of *Dehalococcoides* involved in the dechlorination of various CAHs were also shown to be present in Borehole 2 (Fig. S5 in Supplementary Material) albeit at a moderate level.

Bacteria have been shown to play a major role in hydrocarbon degradation due their ability to utilize the hydrocarbons to satisfy their cell growth and energy needs. According to [41], bioremediation of sites which are severely polluted is achieved by the synergistic interactions of members of the microbial population. It is possible that one type of microorganism removes the toxic metabolites that may inhibit the activities of other microorganisms present in the population. In addition, one microorganism may be able to degrade compounds that another degrades only partially. These findings provided evidence regarding the adaptive abilities of microbial populations residing in contaminated environments. However, to draw similar conclusions from the current study, the microorganism

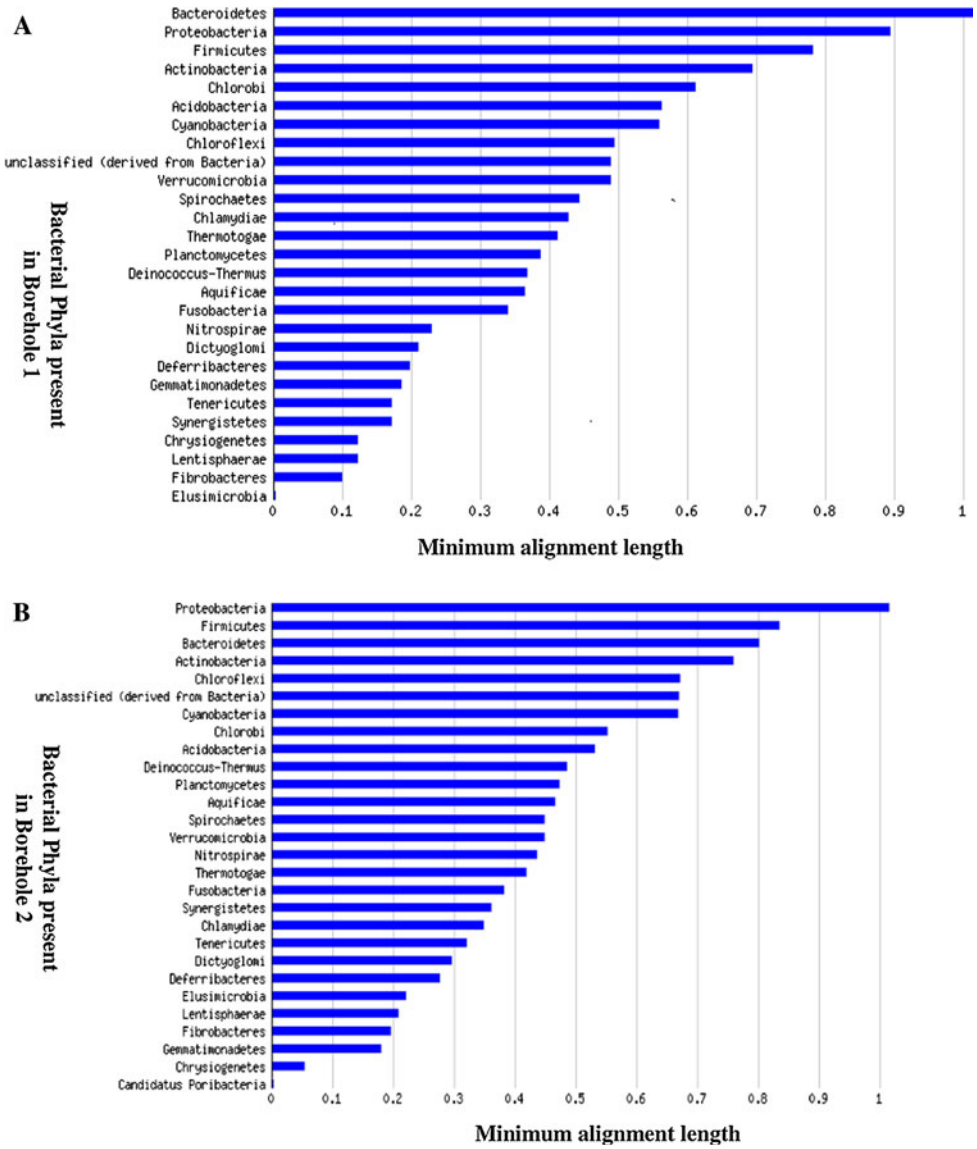


Fig. 1 Taxonomic assignments of the pyrosequencing data using the MG-RAST's underlying SEED database. This data were calculated using a maximum e-value of 1e−0, a minimum identity of 0 %, and a minimum alignment length of 1. The data have been normalized to values between 0 and 1. **a** Borehole 1 shows that the majority of the sequences cluster within the *Bacteriodetes/chlorobi* group to which *Flavobacteria* belong and this microorganism was also shown to be the dominant organism at this site according to the protein recruitment plots and CARMA classification; however, **b** Borehole 2 displayed a majority of SEED hits with *Proteobacteria,* which was comprised mainly of the lineage *Rhodocyclales*, an order to which known hydrocarbon degraders form part of

will need to be isolated from the sample material and their hydrocarbon degradative abilities elucidated.

Metabolic Profile of the Metagenomes

By translating the nucleotide sequences in all six reading frames and mapping the translated sequences to known proteins present in the SEED database, a metabolic profile of the metagenome was generated using the MG-RAST pipeline. This approach has previously been used to analyze the metabolic signatures of a number of metagenomes [38]. The analysis (annotation) was performed by MG-RAST's underlying SEED database [28]. The annotation using SEED occurs through the development of subsystems. Subsystems are referred to as groups of genes that function together, the products of which are involved in a metabolic pathway or make up a cellular structure [10]. The metabolic pathways are classified in a hierarchical structure. At the highest level of organization, the subsystems include both catabolic and anabolic functions and at the lowest levels, the subsystems are specific pathways [33]. A summary of the subsystem categories is shown in (Fig. 2) and all matches to subsystems are provided as supplementary material (Additional Files 3 and 4).

The subsystems present in the sequenced groundwater samples indicated that the pyrosequencing-generated data represented a large percentage of clustering-based subsystems, protein metabolism, carbohydrates, cell wall and capsule, amino acids and derivatives, and respiration. In addition, the metabolic potential that was expected to be present in the contaminated groundwater, namely the metabolism of aromatic compounds, was evident (Fig. 3). Genes involved in the degradation of aromatics (benzene and benzoate), efflux components (CzcABC, CzcD), MarA

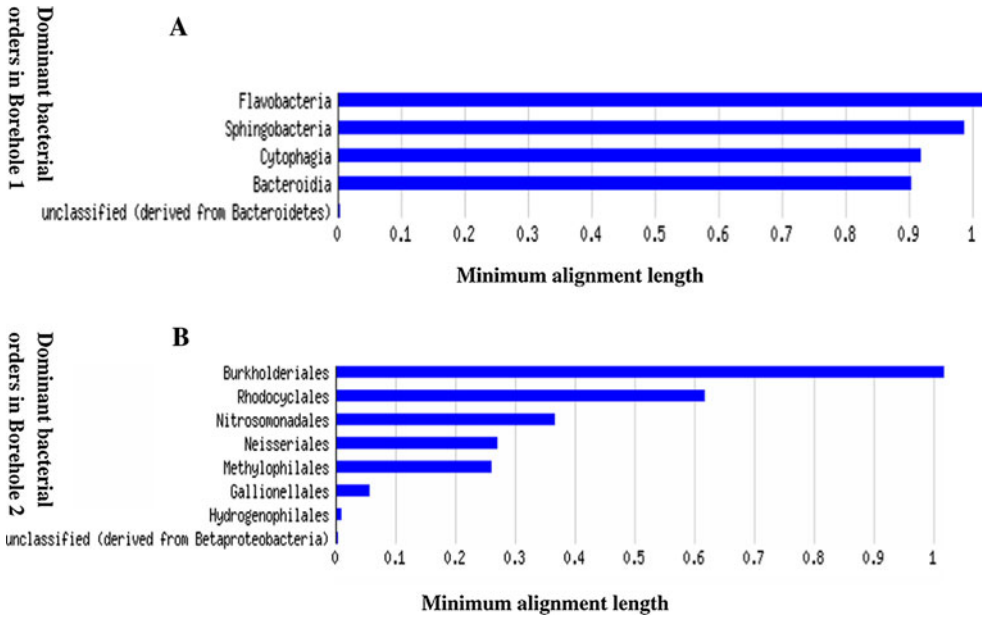


Fig. 2 Taxonomic assignments according to order of the pyrosequencing data using the MG-RAST's-underlying SEED database. This data were calculated using a maximum e-value of 1e−0, a minimum identity of 0 %, and a minimum alignment length of 1. The data have been normalized to values between 0 and 1. **a** Borehole 1 shows that the majority of the sequences cluster within the order *Flavobacteria* and *Sphingobacteria* which house an aromatic degradative potential, **b** Borehole 2 displayed a majority of SEED hits with *Burkholderiales* and *Rhodocyclales (*hydrocarbon degraders). Both these orders contain species that are reported to use hydrocarbons as a sole energy source

family, and MerR were most likely associated with a survival strategy of the microbial communities present at the contaminated sites.

Therefore, the subsystems present in the metagenomes may provide some insight into the microbial ecology of the environment [10].

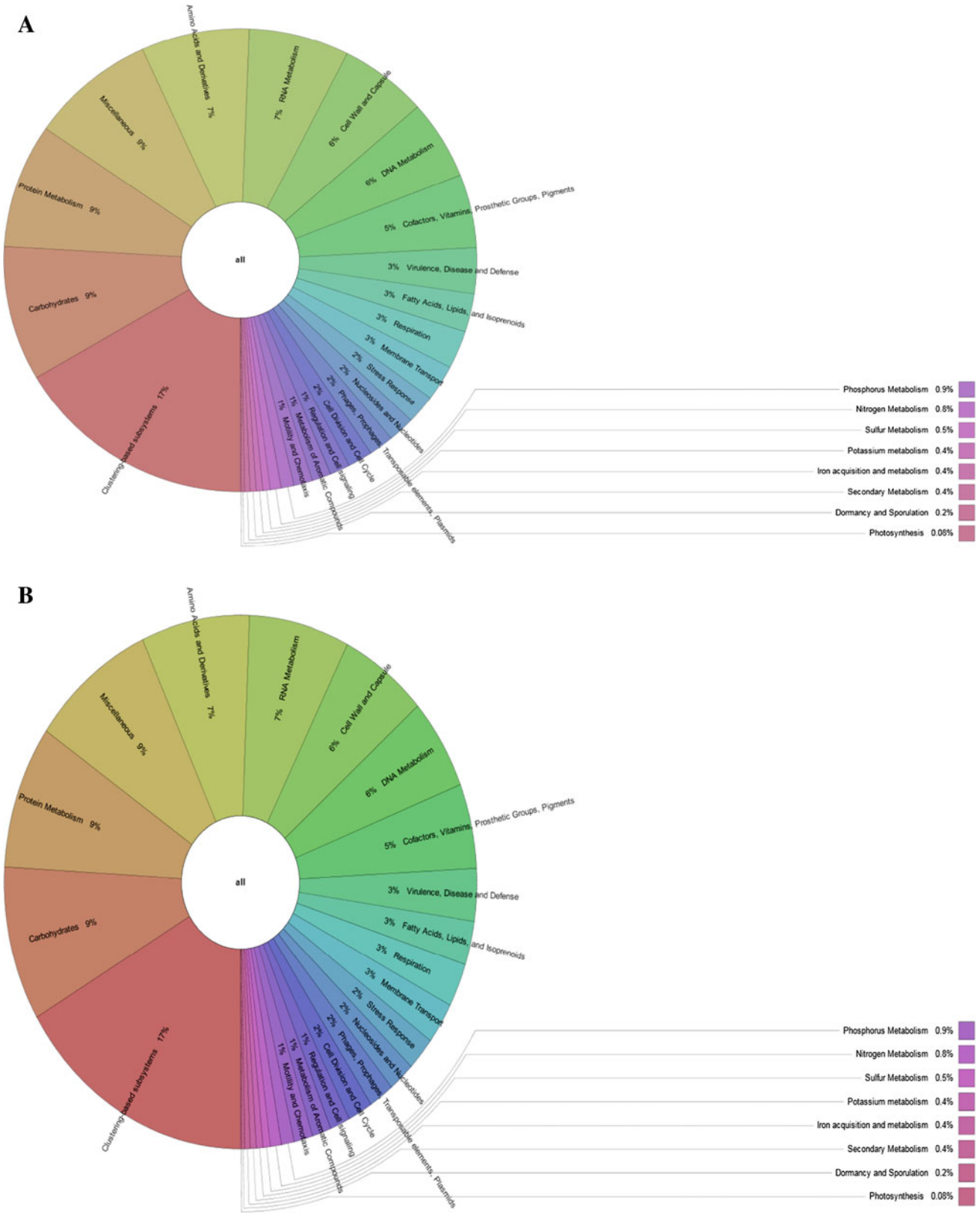
Metabolic Reconstruction of the Metagenomes

Metabolic reconstruction was performed to determine the metabolic capabilities of the microbial community. Sequence coverage of the metagenomes was sufficient for the metabolic reconstruction. Although the reconstruction represents composite cell networks, the information obtained is sufficient to address questions regarding the metabolic potential of the microbial communities and to correlate the data to the contamination profile of the groundwater. Identification of carbon transport systems suggests that the microbial communities persist primarily on monosaccharides and disaccharides. A complete TCA pathway was identified; however, partial pathways for pentose phosphate and the urea cycle were identified. In addition, complete and partial pathways were also identified for the degradation of specific contaminants (e.g., benzene, benzoate, and 1,2-dichloroethane).

Fermentative pyruvate conversion enzymes (pyruvate formate-lyase and pyruvate/ferredoxin oxidoreductase) were identified. It is unclear whether microbial community carries out fermentation to a significant degree as compared to respiration although fermentative species such *Clostridiales* were shown to be present at moderate levels (Fig. S3).

Genes involved in both aerobic and anaerobic respiration were identified in the metagenomes. There was an abundance of genes involved in aerobic respiration, whereby electrons are transferred from hydrogenases to quinones (e.g., ubiquinone) and then to cytochromes as compared to genes involved in anaerobic respiration (nitrate reductase and glutamine synthase-like genes). Aerobic respiration in many bacteria is performed by oxygen reductase members (e.g., cytochrome c oxidase) of the heme-copper oxidoreductase superfamily [7]. These redox-driven proton pumps couple the reduction of O_2 to the translocation of protons across the membrane. Genes/enzymes involved in the biogenesis of cytochrome C oxidases (complex I–IV) were shown to be prevalent in the metagenomes.

The presence of genes involved in nitrogen metabolism (assimilatory nitrate reduction, ammonia assimilation, and allantoin degradation) was identified and this implied that respiration involving inorganic nitrogen sources is present in the groundwater communities. The most abundant nitrogen-cycling genes were shown to be involved in ammonia metabolism, and it appears that the ammonia is being produced through a combination of nitrate/nitrite ammonification and allantoin degradation. Similar observations were



reported for a previous study focusing on freshwater microbialites [4]. The presence of denitrifying enzymes indicates the likely use of nitrate/nitrite as electron acceptors in the absence of oxygen. This is in accordance with the abundance of *Proteobacteria* detected in the groundwater 2(Fig. 1b) since most denitrifying bacteria belong to

◀ **Fig. 3** Subsystems present in the groundwater samples following normalization to reduce redundancies in the sequence data as conducted by MG-RASTs annotation. For both sequenced sites, a majority of the sequences was classified as clustering-based systems and carbohydrate metabolism. Despite the identification of composite cell networks for both sequenced samples, a category denoted metabolism of aromatic compound was shown to be present. The Borehole 2 (**b**) metagenome contained a larger percentage of hydrocarbon degradative genes when compared to the Borehole 1 metagenome (**a**) and this is also linked to the observed degradation profiles (obtained from VOC and GC analysis) of these compounds within the sampled sites

Proteobacteria, which are known to be metabolically versatile facultative aerobic bacteria [6].

Among the major anthropogenic pollutants, chlorinated aliphatics and aromatics are of great concern in environments such as groundwater and sediments where they tend to accumulate [26]. The duration, cost, and uncertain success of pumping out dissolved contamination is a daunting obstacle both to the protection of groundwater and to the recycling and re-development of industrially contaminated land [12]. Their environmental persistence, toxicity, and/or carcinogenicity and potential for bioaccumulation in food chains are of serious environmental concern [15]. Degradation of contaminants requires specific specialized multistep pathways [40]. Microorganisms have been shown to evolve an extensive range of enzymes and pathways that enable them to degrade a wide array of xenobiotics [22].

Chlorinated compounds are probably the most important class of environmental pollutants and include numerous halogenated compounds [27]. Chlorinated ethenes can be biologically degraded to less-chlorinated and, in some cases, completely dechlorinated to the non-toxic end product ethane [17]. As mentioned previously, strains of *Dehalococcoides*, belonging to the phylum *Chloroflexi* were reported to be associated with the degradation of chlorinated hydrocarbons [27]. This genus was shown to be present at a moderate level in Borehole 2. In addition, in a previous study conducted by our research group (data not published), the degradation of CAHs was assessed by gas chromatography. The results obtained indicated that PCE, TCE, DCE, and DCA were readily degraded by the microbial communities present at the contaminated sites. However, the higher concentration of CAHs observed in Borehole 1 as compared to Borehole 2 could be attributed to the very low levels of the phylum *Chloroflexi*, being present. This phylum includes strains involved in CAH degradation (Fig. S2).

In this study, several putative degradative pathways for specific compounds were determined (e.g., butanol [butyryl-CoA dehydrogenase] and 1,2-DCA).

As shown in Table 1, a higher degradation potential for 1,2-DCA, the degradation product of TCE, resulting in an accumulation of VC was observed at Borehole 2 as compared to Borehole 1. However, we were unable to identify a complete CAH degradative pathway.

Table 3 Collection of gene products associated with benzoate/benzene catabolism and degradation accessed from both metagenomes (Borehole 1 and 2)

Gene products	Function
2-Ketocyclohexanecarboxyl-CoA hydrolase (EC 4.1.3.36)	Anaerobic benzene ring biodegradation
Beta-ketoadipate enol-lactone hydrolase (EC 3.1.1.24)	Catechol branch of beta-ketoadipate pathway
3-Oxoadipate CoA-transferase subunit A (EC 2.8.3.6)	Catechol branch of beta-ketoadipate pathway
Mandelate racemase/muconate lactonizing enzyme family protein	Catechol branch of beta-ketoadipate pathway
4-Oxalocrotonate decarboxylase (EC 4.1.1.77)	Central meta-cleavage pathway of aromatic compound degradation
Phenol hydroxylase large subunit	Central meta-cleavage pathway of aromatic compound degradation
2-Hydroxymuconic semialdehyde hydrolase (EC 3.7.1.9)	Central meta-cleavage pathway of aromatic compound degradation
3,4-Dihydroxyphenylacetate 2,3-dioxygenase (EC 1.13.11.15)	Central meta-cleavage pathway of aromatic compound degradation
5-Carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (EC 1.2.1.60)	Central meta-cleavage pathway of aromatic compound degradation
2,4-Dihydroxyhept-2-ene-1,7-dioic acid aldolase (EC 4.1.2.-)	Central meta-cleavage pathway of aromatic compound degradation
3-Oxoadipate CoA-transferase subunit B (EC 2.8.3.6)	Benzoate catabolism
Benzoate 1,2-dioxygenase (EC 1.14.12.10)	Benzoate catabolism
Benzoate dioxygenase, ferredoxin reductase component	Benzoate degradation
Benzoate catabolic operon transcription regulator	Benzoate degradation
Aromatic hydrocarbon utilization transcriptional regulator CatR (LysR family)	Benzoate degradation
BenABC operon transcriptional activator BenR	Benzoate degradation
Benzoate transport protein	Benzoate degradation

In addition to carbohydrates, aromatic compounds represent the second most abundant class of natural products and are efficiently used as growth substrates by microorganisms [11]. Degradation of aromatic compounds is mainly performed by microorganisms and occurs in oxic as well as anoxic environments [20]. Microorganisms containing various dioxygenases capable of cleaving aromatic compounds (e.g., benzoate and benzene) have been reported [20]. During degradation, the aromatic compounds are invariably converged by dioxygenases into more reactive dihydroxylated intermediates, e.g., catechol [20]. Catechol is then degraded by either catechol 1,2-dioxygenase (*ortho* cleavage pathway) or catechol 2,3-dioxygenase (*meta* cleavage pathway) [39]. The intermediates of benzoate and benzene degradation pathways eventually lead to the TCA cycle. Genes involved in benzoate catabolism were also identified (Table 3). Taking into considerable the data on the recruitment plots for Borehole 2, it is not uncommon to identify genes involved in aerobic benzene and benzoate degradation since the ability of *Azoarcus* to degrade aromatics aerobically has been well documented. Similar observations were made in the study of Hemme et al. [14], in which no complete pathways for aromatic compounds were elucidated from a stressed groundwater community. In addition, by generation of a heatmap using the MG-RAST pipeline, we compared the subsystems data (normalized data) between the two boreholes. The heatmap/dendrogram is a tool that allows an enormous amount of information (e.g., the abundance values of thousands of functional roles across dozens of metagenomic samples) to be presented in a visual form that is amenable to human interpretation. Dendrograms are trees that indicate how similar/dissimilar a group of vectors (list of values, like the abundance counts from a single metagenome) are to each other. Vectors in a dendrogram are usually ordered with respect to their level of similarity: similar vectors are placed next to each other and more distantly related vectors are placed further apart. The MG-RAST heatmap/dendrogram has two dendrograms, one indicating the similarity/dissimilarity among metagenomic samples (*x*-axis dendrogram) and another to indicate the similarity/dissimilarity among categories (e.g., functional roles; the *y*-axis dendrogram. A distance metric (euclidean distance) was used to determine the similarity/dissimilarity between every possible pair of sample abundance profiles. The resulting distance matrix was used with a clustering algorithm. Each square in the heatmap dendrogram represents the abundance level (as MG-RAST normalized values) of a single category in a single sample. Using the subsystems data from the Borehole 1 and 2, a heatmap was generated (Fig. 4). According to the dendrogram, there were no major differences in the abundance ranking observed for the subsystems of interest in this study, such as the metabolism of aromatic compounds, virulence, and stress response in the individual metagenomes. These data were consistent with the observed subsystem categories which have been represented as percentages in Fig. 3.

From the literature, it was observed that numerous metagenome-sequencing studies from hydrocarbon-contaminated sites [1, 5, 8] focused on the amplification of the 16S rRNA gene or the detection of a single resistance gene rather than analyzing the sequence data of the entire microbial community. Therefore, our basis for comparisons with published 454 metagenomic data was limited. However, with the available whole genome sequences present in the MG-RAST database, we performed a principal component analysis (PCA) which is part of the MG-RAST pipeline. PCA allows for samples which exhibit similar abundance profiles (taxonomic or functional) to be grouped together. This analysis was performed on the normalized data. According to the analysis (Fig. 5), the functional categories for the studied metagenomes showed dissimilarity to other aquatic metagenomes (coastal surface waters, marine water metagenomes and freshwater metagenomes). Based on this grouping profile, it is evident that the sequences present in the microbial communities of the studied metagenomes are dissimilar to other aquatic environments, thereby meriting the uniqueness of the sequences and processes of the studied microbial communities.

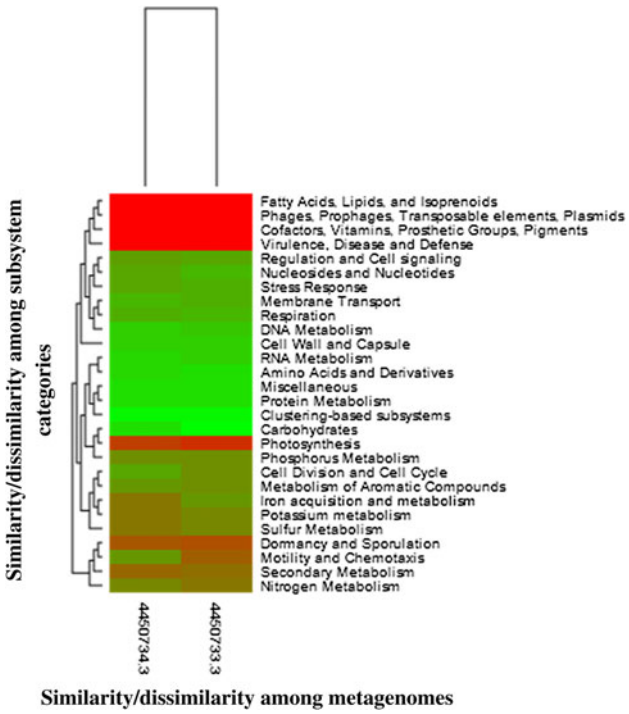


Fig. 4 Heatmap generated for Borehole 1 and 2 using the MG-RAST pipeline. The cluster analysis represents the SEED subsystems. Analysis was based on the relative abundances of the non-redundant protein dataset within each metagenome. The *each square* in the heatmap dendrogram represents the abundance level with 0 being the least abundance (*red*) and 1 being the most abundant (*bright green*)

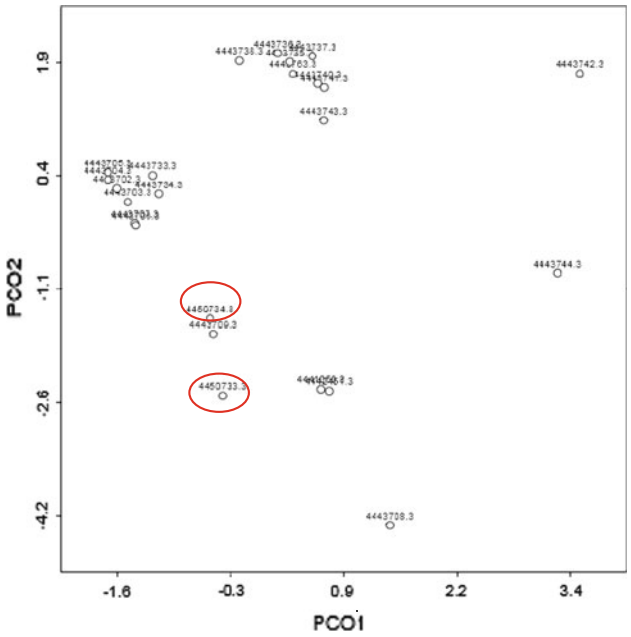


Fig. 5 PCA (principal component analysis) of the studied metagenomes (indicated by the *red circle*) together with whole genome sequence data which are available in the MG-RAST public and project databases. Comparisons were with sequence data available from other aquatic sequencing based projects. PCA allows for samples which exhibit similar abundance profiles (taxonomic or functional) to be grouped together. From the generated data figure, it is evident that the sequence data contained in the studied metagenomes is quite dissimilar to the sequence data present in other aquatic environments based on functional (SEED) abundance ranks

From our data, it was evident that there is a higher percentage of stress response genes (2 %) and virulence genes (3 and 4 %, respectively) in both boreholes when compared to genes involved in the metabolism of aromatic compounds (0.8 and 1 %, respectively); therefore, we suggest that the communities may be compensated for by a more general stress response systems such as metal efflux proteins, acid resistance mechanisms, and the presence of virulence determinants (resistance to antibiotics and toxic compounds) (Additional Files 3 and 4). A similar scenario was observed in data published on stressed groundwater metagenomes investigated by Hemme [14] and Smith [37].

A complement of heavy metal-resistance genes Cd^{2+}/Zn^{2+}/Co^{2+} efflux components (CzcABC, CzcD) divalent cation transporters and mercuric resistance genes has also been identified in the groundwater communities. Genes involved in heavy metal efflux have been identified in *D. aromaticus*, the sequences of which were present in abundance in Borehole 2 [29]. In addition, mercury-resistant genes (*mer*) are often located on transposons or conjugative plasmids [24]. In this study, the broad host range conjugative plasmid (pSB102) conferring mercury resistance was identified only in the Borehole 2 metagenome.

The results obtained revealed insights into microbial community diversity, structure, and function as well as suggested possible processes by which microbial communities adapt to environmental contamination. It is understood that conclusions on the functional role of the microorganisms inhabiting contaminated sites cannot be made. This is because the amount of sequence information generated for individual habitats is still relatively limited [34]. The results obtained in this study indicated that the groundwater communities were able to metabolize a host of complex compounds; however, additional experiments are vital to establish a link between the occurrences of catabolic genes that were identified by the sequencing with in vivo catabolism of the compounds.

Adaptation of microbial communities to environmental stresses is a key issue in ecology [14]. The pyrosequencing analysis has revealed that the microbial communities residing at the contaminated sites are well adapted to the geochemical conditions of the groundwater. The identification of degradative genes and detoxification pathways in these microorganisms have a potential application for bioremediation. However, their full potential is yet to be exploited.

Conclusions

The focus of this paper was to provide genomic information of hydrocarbon contaminated boreholes in Durban, South Africa. This study was the first to employ high-throughput sequencing analysis of the respective sites, and has therefore served as a foundation for numerous future studies which will include bioremediation. Analysis of the taxonomic composition and metabolic potential revealed that the communities are dominated by the *Bacteriodetes* and *Proteobacteria*, both groups encompassing aromatic degraders. Although the metabolic reconstruction represented composite cell networks, the information obtained was sufficient to address questions regarding the metabolic potential of the microbial community and to correlate the data to the contamination profile of the Boreholes. The metabolic potential that was expected to be present in the contaminated metagenomes, including metabolism of aromatic compounds (benzene and benzoate degradative pathways), and a general stress response system (virulence, heavy metal resistance, efflux proteins) were evident. Therefore, this study provided insight regarding the growth and survival strategies for microbial communities inhabiting contaminated environments.

Acknowledgments The authors wish to thank the National Research Foundation for financial support and TIA (Dr James Sakwa and colleagues) for the pyrosequencing, Dr Rehana Shaik and Dr Algasan Govender for the sample collection and processing.

References

1. Aburto, A., Fahy, A., Coulon, F., Lethbridge, G., Timmis, K., et al. (2009). Mixed aerobic and anaerobic microbial communities in benzene-contaminated groundwater. *Journal of Applied Microbiology, 106*, 317–328.
2. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389–3402.
3. Beller, H. R., Kane, S. R., Legler, T. C., & Alvarez, P. J. J. (2002). A real-time polymerase chain reaction method for monitoring anaerobic, hydrocarbon-degrading bacteria based on a catabolic gene. *Environmental Science and Technology, 36*, 3977–3984.
4. Breitbart, M., Hoare, A., Nitti, A., Siefert, J., Haynes, M., et al. (2009). Metagenomic and stable isotopic analyses of modern freshwater microbialites in Cuatro Cienegas, Mexico. *Environmental Microbiology, 11*, 16–34.
5. Brennerova, M. V., Josefiova, J., Brenner, V., Pieper, D. H., & Junca, H. (2009). Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environmental Microbiology, 11*, 2216–2227.
6. Bru, D., Sarr, A., & Philippot, L. (2007). Relative abundances of proteobacterial membrane-bound and periplasmic nitrate reductases in selected environments. *Applied and Environmental Microbiology, 73*, 5971–5974.
7. Chang, H. Y., Hemp, J., Chen, Y., Fee, J. A., & Gennis, R. B. (2009). The cytochrome ba3 oxygen reductase from Thermus thermophilus uses a single input channel for proton delivery to the active site and for proton pumping. *Proceedings of the National Academy of Sciences, 106*, 16169–16173.
8. Chikere, C. B., Okpokwasili, G. C., & Chikere, B. O. (2011). Monitoring of microbial hydrocarbon remediation in the soil. 3 *Biotech*, 1–22.
9. Davis, G. B., Patterson, B. M., & Johnston, C. D. (2009). Aerobic bioremediation of 1,2 dichloroethane and vinyl chloride at field scale. *Journal of Contaminant Hydrology, 107*, 91–100.
10. Edwards, R., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., et al. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics, 7*, 57.
11. Gescher, J., Eisenreich, W., Wörth, J., Bacher, A., & Fuchs, G. (2005). Aerobic benzoyl-CoA catabolic pathway in *Azoarcus evansii*: Studies on the non-oxygenolytic ring cleavage enzyme. *Molecular Microbiology, 56*, 1586–1600.
12. Gualandi, G., Frascari, D., Pinelli, D., & Nocentini, M. (2007). Growth of chlorinated solvent-degrading consortia in fed-batch bioreactors and development of a double-substrate high-performing microbial inoculum. *Engineering in Life Sciences, 7*, 217–228.
13. Hemalatha, S., & Veeramanikandan, P. (2011). Characterization of aromatic hydrocarbon degrading bacteria from petroleum contaminated sites. *Journal of Environmental Protection, 2*, 243–254.
14. Hemme, C. L., Deng, Y., Gentry, T. J., Fields, M. W., Wu, L., et al. (2010). Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *The ISME Journal, 4*, 660–672.
15. Horvath, R. S. (1972). Microbial co-metabolism and the degradation of organic compounds in nature. *Bacteriological Reviews, 36*, 146.
16. Huson, D., Richter, D., Mitra, S., Auch, A., & Schuster, S. (2009). Methods for comparative metagenomics. *BMC Bioinformatics, 10*, S12.
17. Johnson, D. R., Lee, P. K. H., Holmes, V. F., Fortin, A. C., & Alvarez-Cohen, L. (2005). Transcriptional expression of the tceA gene in a Dehalococcoides-containing microbial enrichment. *Applied and Environmental Microbiology, 71*, 7145–7151.
18. Jurelevicius, D., Alvarez, V. M., Peixoto, R., Rosado, A. S., & Seldin, L. (2012). Bacterial polycyclic aromatic hydrocarbon ring-hydroxylating dioxygenases (PAH-RHD) encoding genes in different soils from King George Bay, Antarctic Peninsula. *Applied Soil Ecology, 55*, 1–9.
19. Kennedy, J., Flemer, B., Jackson, S. A., Lejon, D. P. H., Morrissey, J. P., et al. (2010). Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. *Marine Drugs, 8*, 608–628.
20. Kim, S. J., Kweon, O., & Cerniglia, C. E. (2009). Proteomic applications to elucidate bacterial aromatic hydrocarbon metabolic pathways. *Current Opinion in Microbiology, 12*, 301–309.
21. Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., et al. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research, 36*, 2230–2239.
22. Krooneman, J., Sliekers, A. O., Pedro Gomes, T. M., Forney, L. J., & Gottschal, J. C. (2000). Characterization of 3-chlorobenzoate degrading aerobic bacteria isolated under various environmental conditions. *FEMS Microbiology Ecology, 32*, 53–59.
23. Mao, J., Luo, Y., Teng, Y., & Li, Z. (2012). Bioremediation of polycyclic aromatic hydrocarbon-contaminated soil by a bacterial consortium and associated microbial community changes. *International Biodeterioration and Biodegradation, 70*, 141–147.
24. Margesin, R., & Schinner, F. (2001). Biodegradation and bioremediation of hydrocarbons in extreme environments. *Applied Microbiology and Biotechnology, 56*, 650–663.
25. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*, 376–380.
26. Marzorati, M., Balloi, A., De Ferra, F., Corallo, L., Carpani, G., et al. (2010). Bacterial diversity and reductive dehalogenase redundancy in a 1,2-dichloroethane-degrading bacterial consortium enriched from a contaminated aquifer. *Microbial Cell Factories, 9*, 12.
27. Marzorati, M., De Ferra, F., Van Raemdonck, H., Borin, S., Allifranchini, E., et al. (2007). A novel reductive dehalogenase, identified in a contaminated groundwater enrichment culture and in Desulfitobacterium dichloroeliminans strain DCA1, is linked to dehalogenation of 1, 2-dichloroethane. *Applied and Environmental Microbiology, 73*, 2990–2999.
28. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., et al. (2008). The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*, 386.
29. Nemergut, D. R., Robeson, M. S., Kysela, R. F., Martin, A. P., Schmidt, S. K., et al. (2008). Insights and inferences about integron evolution from genomic data. *BMC Genomics, 9*, 261.
30. Palmer, J., Oliver, A., & Cameron-Clarke, I. (2001). Containment of groundwater contaminated with chlorinated hydrocarbons, Umbogintwini Industrial Complex, South Africa. *Groundwater quality: Natural and enhanced restoration of groundwater pollution*, 579–584.
31. Rees, H. C., Oswald, S. E., Banwart, S. A., Pickup, R. W., & Lerner, D. N. (2007). Biodegradation processes in a laboratory-scale groundwater contaminant plume assessed by fluorescence imaging and microbial analysis. *Applied and Environmental Microbiology, 73*, 3865–3876.
32. Salinero, K. K., Keller, K., Feil, W. S., Feil, H., Trong, S., et al. (2009). Metabolic analysis of the soil microbe *Dechloromonas aromatica* str. RCB: Indications of a surprisingly complex lifestyle and cryptic anaerobic pathways for aromatic degradation. *BMC Genomics, 10*, 351.
33. Sanapareddy, N., Hamp, T. J., Gonzalez, L. C., Hilger, H. A., Fodor, A. A., et al. (2009). Molecular diversity of a North

Carolina wastewater treatment plant as revealed by pyrosequencing. *Applied and Environmental Microbiology, 75*, 1688–1696.

34. Schmeisser, C., Stöckigt, C., Raasch, C., Wingender, J., Timmis, K., et al. (2003). Metagenome survey of biofilms in drinking-water networks. *Applied and Environmental Microbiology, 69*, 7298–7309.
35. Shen, Y., Stehmeier, L. G., & Voordouw, G. (1998). Identification of hydrocarbon-degrading bacteria in soil by reverse sample genome probing. *Applied and Environmental Microbiology, 64*, 637–645.
36. Simon, C., Wiezer, A., Strittmatter, A. W., & Daniel, R. (2009). Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome. *Applied and Environmental Microbiology, 75*, 7519–7526.
37. Smith, R. J., Jeffries, T. C., Roudnew, B., Fitch, A. J., Seymour, J. R., et al. (2012). Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. *Environmental Microbiology, 14*, 240–253.
38. Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., et al. (2005). Comparative metagenomics of microbial communities. *Science, 308*, 554–557.
39. Vaillancourt, F. H., Bolin, J. T., & Eltis, L. D. (2006). The ins and outs of ring-cleaving dioxygenases. *Critical Reviews in Biochemistry and Molecular Biology, 41*, 241–267.
40. Weightman, A. J., & Marchesi, J. R. (2003). Comparing the dehalogenase gene pool in cultivated α-halocarboxylic acid-degrading bacteria with the environmental metagene pool. *Applied and Environmental Microbiology, 69*, 4375–4382.
41. Yusoff, W. M. W. (2008). Development of three bacteria consortium for the bioremediation of Crude petroleum-oil in contaminated water. *OnLine Journal of Biological Sciences, 8*, 73–79.
42. Zaar, A., Gescher, J., Eisenreich, W., Bacher, A., & Fuchs, G. (2004). New enzymes involved in aerobic benzoate metabolism in *Azoarcus evansii*. *Molecular Microbiology, 54*, 223–238.