Analysis of viral and bacterial communities in groundwater associated with contaminated land

Costeira, R., Doherty, R., Allen, C. C. R., Larkin, M. J., & Kulakov, L. A. (2019). Analysis of viral and bacterial communities in groundwater associated with contaminated land. *Science of the Total Environment*, *656*, 1413-1426. https://doi.org/10.1016/j.scitotenv.2018.11.429

Analysis of Viral and Bacterial Communities in Groundwater Associated with Contaminated Land

Ricardo Costeira[1], Rory Doherty[2], Christopher CR Allen[1,3], Michael J Larkin[1], Leonid A Kulakov[1*]

[1]School of Biological Sciences, Queen's University Belfast, UK

[2]School of the Natural and Built Environment, Queen's University Belfast, UK

[3]Institute for Global Food Security, Queen's University Belfast, UK

*Corresponding author: School of Biological Sciences, The Queen's University of Belfast, 97 Lisburn Road, Belfast, Northern Ireland BT9 7BL, UK.

E-mail: l.kulakov@qub.ac.uk

Highlights

- Bacteriophages are considered to be key entities of various environments
- Groundwater microbial communities were studied using molecular biology approaches
- Phage and bacterial diversities were correlated with contamination and pH
- Viruses of degraders were identified and phage-bacterial associations described
- A total environmental community approach provides valuable insights towards bioremediation

Abstract

Bacteriophages play a role in the diversification and production of bacteria whithin complex communities of microbes, and are thought to influence local bacterial degrader capacities. Here, we report a joint metagenomic characterization of the bacterial and viral communities of groundwater associated with a contaminant plume, and examine the extent of their interactions. Over a year, 14 metagenomes and viromes were collected at different locations from an old gasworks site and sequenced using Illumina next generation sequencing technologies. We show that the viral community diversity mirrored the bacterial diversity found. Bacterial degraders were abundant at the site (*e.g. Thermoanaerobacteriaceae*, *Caulobacteraceae*) as were virotypes of degraders (*e.g. Thermoanaerobacterium* phage THSA-485A, *Caulobacter* phage CcrColossus). Host assignment of the viral communities revealed that interactions were limited to few classes of bacteria (*e.g.* Clostridia and Proteobacteria) and that these were discrete across the site. Putative viral generalists infecting multiple species of degraders were identified. Overall, findings reported support the need of phage research while designing bioremediation strategies.

Keywords

bacteria, bacteriophages, water resources, gasworks, bioremediation, metagenomics

1. Introduction

Since the advent of industrialization, a range of anthropogenic activities have led to an abundance of contaminants in the environment. As of now, at least 127,000 contaminated sites have been identified in Europe and more than 342,000 sites have been extrapolated to be polluted in the whole continent (Panagos et al., 2013). Groundwater contamination may occur from various point sources due to accidental spills, landfills, oil pipelines and land misuse, or from widespread application of contaminants due to agriculture and sewage treatments (Brandon, 2013; Meckenstock et al., 2015). Groundwater contamination not only leads to the depletion of pristine fresh water reserves, but also impacts the total environment and poses serious risks to human health (Danielopol et al., 2003). The management and remediation of contaminated sites in Europe is thought to cost around 6 billion Euros annually and bioremediation strategies have gained wide interest as an environmentally friendly and cost-effective way to remediate groundwater and sediment (Majone et al., 2015; Panagos et al., 2013).

Bioremediation strategies are based on the exploitation of the extensive metabolic versatility of microbes, particularly bacteria, to clean-up environmental contaminants that function as nutrient or energy sources for bacterial cells (Aracic et al., 2015). Different strategies of bioremediation exist. Particularly, metagenomic-based bioremediation approaches provide a comprehensive and detailed knowledge of endemic uncultured bacterial populations and allow scientists to describe, exploit and monitor the local biodegradative capacity of the local microbial communities (Devarapalli and Kumavath, 2015).

In 2015, Meckenstock et al. suggested that bacteriophages, *i. e.* viruses that infect bacteria, may play important roles in bioremediation processes. Bacteriophages (or simply, phages) are the most abundant and ubiquitous biological entities known to mankind, with an excess of 1e31 viral-like particles (VLPs) estimated to exist globally (Clokie et al., 2011; Rohwer et al., 2009). A constant ratio of 3-10 VLPs per bacterium has been found in aquatic ecosystems (Wommack and Colwell, 2000). Bacteriophages require obligatory host infection to complete their life cycles (Clokie et al., 2011) and, due to this, dynamic interactions between phages and bacteria are observable in nature, often determining the success of distinct bacterial populations within complex communities of microbes (Clokie et al., 2011). Viral-bacterial interactions can range from predatory to mutualistic (Weinbauer and Rassoulzadegan, 2004). During lytic infections, phages keep in check the dominant bacteria, allowing the co-existence of other bacterial species, (known as the "kill-the-winner" hypothesis) and contributing to the Earth's carbon cycling by the release of organic matter from lysed cells (also known as viral shunt) (Rohwer et al., 2009; Weinbauer and Rassoulzadegan, 2004). Phages have also been described as important for genetic diversity by mediating the horizontal transfer of genes within microbial communities through generalized and specialized transduction (Canchaya et al., 2003). Moreover, the occurrence of auxiliary metabolic genes within phage genomes can reprogram the metabolism of bacterial cells and increase the fitness of bacterial populations (Breitbart, 2012). This may lead to the reshaping and diversification of prokaryotic degrader communities and, thus, influence *in situ* biodegradation rates (Meckenstock et al., 2015).

Up till now, only few studies have been published on viruses of groundwater (Eydal and Jägevall, 2009; Kyle and Eydal, 2008; Pan et al., 2017; Smith et al., 2013) and, to our knowledge, no metagenomic study of viral diversity in this environment has been reported. Moreover, there have only been limited studies of viral diversity and viral roles in polluted waters (Marie and Lin, 2017; O'Brien et al., 2017).

Here, we present a metagenomic characterization of viral communities around a contaminated groundwater plume, and study the dynamics of their interactions with local populations of bacteria. A year-long metagenomic study was carried out at an old gasworks site in Northern Ireland. The site suffers from typical hydrocarbon pollution and has a heterogeneous contaminant distribution. Bacterial and viral community structures and bacteriophage host populations were characterized at different locations at the site during the sampling period. The impact of our findings on natural attenuation and design of bioremediation strategies was hypothesized.

2. Materials and Methods

2.1 Site of study and sampling design

The gasworks site studied here operated for 163 years (1822-1985) in an urban area of Northern Ireland and has undergone remediation by excavation over several phases during the mid-1990s. Its land has been repurposed since. Permit to access and sample the site was given by the local council. Six sampling stations were selected and their groundwater chemistry is in Supplementary Data A. Three of the sampling stations selected had access to hydrocarbon-contaminated groundwater ("C") and three of the sampling stations had access to groundwater showing no previous traces of hydrocarbon contamination. Samples from these stations were hereby referred to as non-contaminated groundwater samples ("NC").

Over one year (May 2016-May 2017), groundwater was collected every three months from two sampling stations (C1 and NC1) in order to characterize temporal changes within complex communities of microbes at the site. During this period, additional groundwater sampling was done at other stations at the site in order to evaluate spatial variations of bacterial and viral community structures and interactions. The location of the sampling stations used in this study and the timeline of sampling is presented in Figure 1.

At each sampling event, a minimum of 15 L of groundwater were collected with a bailer. Different bailers were used at each sampling station. Any stagnant groundwater in the boreholes was purged before sampling. During sampling, groundwater collected with bailers was mixed in large sterile containers and kept at 4 ºC until processing. Processing occurred within 24 hours of sample collection.

2.2 Sample processing and DNA Isolation

Large particles of sediment were removed from groundwater samples using sterile GF/A glass microfiber filters (Whatman/GE Healthcare, UK). Isolation of total metagenomic DNA and viral metagenomic DNA followed.

Five litres of groundwater per sample were used for isolation of total metagenomic DNA. Microbial cells were recovered using 0.45 µm mixed cellulose ester membrane filters (Whatman/GE Healthcare, UK). Total metagenomic DNA was extracted and purified using the PowerWater DNA Isolation kit (MO BIO, USA).

Ten litres of groundwater per sample were used for isolation of viral metagenomic DNA. Isolation and concentration of VLPs from groundwater samples was done as described by Skvortsov et al. (2016) and Thurber et al. (2009). Briefly, bacterial cells were removed, and VLPs were concentrated to a final volume of 35-50 mL. Epifluorescence microscopy was performed at every step to monitor the absence of bacterial contamination in the final concentrates and DNAse I reactions were performed to further ensure that VLP concentrates were free of any contamination with environmental DNA. For DNA isolation, formamide/CTAB extractions followed by phenol/chloroform purifications were performed (Thurber, 2011).

Quantification of total metagenomic and viral metagenomic DNA was performed with a Quantus fluorometer using the QuantiFluor dsDNA system (Promega, USA).

2.3 Next Generation Sequencing

16S rRNA amplicon sequencing. Total metagenomic DNA from groundwater samples was used for amplification and sequencing of bacterial 16S rRNA genes at Molecular Research LP (USA). Amplicons of the 16S rRNA gene were generated using primers targeting the V4 variable region (515/806) (Soergel et al., 2012) with a barcode on the forward primer. A 30 cycle PCR reaction was performed using the HotStarTaq Plus Master Mix Kit (Qiagen, USA). Briefly, DNA was denatured at 95°C for 5min, amplified with 28 cycles of denaturation at 94°C for 30s, annealing at 53°C for 40s and extension at 72°C for 1min, and finally extended for 5min at 72°C. PCR products were purified with calibrated AMPure XP Beads (Beckman Coulter Inc, USA) and DNA libraries were prepared using an Illumina TruSeq DNA library protocol (Illumina Inc, USA). Sequencing of 2 x 300 bp (PE) amplicon libraries was performed on the Illumina MiSeq System using MiSeq Reagent Kit v3 chemistry (Illumina Inc, USA).

Shotgun sequencing. Total metagenomic DNA and viral metagenomic DNA isolated from groundwater samples were used for whole metagenome shotgun sequencing at the Centre for Genomic Research of the University of Liverpool (UK). Prior to library preparation, DNA was bead-purified and quality-controlled by capillary electrophoresis with a Fragment Analyzer (Advanced Analytical Technologies Inc, USA). The Nextera XT DNA Library Prep Kit (Illumina Inc, USA) was used for metagenomic library preparation. DNA libraries of 2 x 150 bp (PE) were sequenced with Illumina HiSeq 2500/HiSeq 4000 Systems using the latest SBS chemistry (Illumina Inc, USA).

2.4 Bioinformatic analysis

Bacterial community diversity analysis. 16S rRNA gene amplicon read pairs were trimmed (Q25) on both ends and merged at the sequencing facility. Quantitative sequencing analysis was carried out using QIIME 1.9.1 (Caporaso et al., 2010a). Sequences were demultiplexed and barcodes were removed. Clustering of sequences into OTUs was performed using open-reference OTU picking based on 97% similarity with USEARCH v6.1.544 (Edgar, 2010). Sequence alignment was done with PyNAST 1.0 (Caporaso et al., 2010b) and taxonomy assignment was done using the most recent Greengenes reference database (August 2013) (DeSantis et al., 2006) with the UCLUST algorithm (Edgar, 2010). Core diversity analysis was performed after sample BIOM table rarefaction for sampling depth normalization. OTUs were used for estimation of sample diversity. Sample diversity analysis and sample cluster analysis were performed using the vegan v2.5-2 R package (Oksanen et al., 2018). Bray-Curtis as was used as dissimilarity method.

Viral community diversity analysis. Virome shotgun reads were trimmed and quality filtered at the sequencing facility using Cutadapt (Martin, 2011) and Sickle v1.200 (Joshi and Fass, 2011). Read pairs were quality controlled using FastQC (Andrews, 2010) and merged using PEAR v0.9.8 (p-value = 0.01, min. overlap size = 10 bp, min length = 50 bp. Q = 33) (Zhang et al., 2014). Processed reads were assembled into contigs using metaSPAdes (SPAdes v3.9.0; k-mer sizes = 21, 33, 55,77 bp) (Nurk et al., 2017). Metagenome assemblies were quality assessed using MetaQUAST (Mikheenko et al., 2016). Identification of ORFs in contigs was done with Prodigal v2.6 (-g 11 -p meta) (Hyatt et al., 2010) and proteins files were blasted against the Viral RefSeq database (accessed 11 January 2017) using DIAMOND v0.8.34.96 (BLASTp, e-value = 0.001) (Buchfink et al., 2014). DIAMOND blast files were imported into MEGAN6 (Huson et al., 2016) and taxonomic analysis was performed using the LCA algorithm after sample rarefaction (min. percent identity score = 50, top percent hits = 10, min. taxon assignment percent = 0.01). Sample cluster analysis was performed using annotated virotypes and using Bray-Curtis as dissimilarity method. Sample diversity analysis was performed at different taxonomic levels.

Viral-bacterial interaction analysis. For host assignment of bacteriophages, contigs ≥ 2.5 kb were considered. Taxonomic assignment of larger contigs was performed with CAT v1.0 (Cambuy et al., 2016) using annotation results of protein alignments against the Viral RefSeq database. Contigs assigned under viral domain (VCs) were kept and unassigned contigs were filtered out. Host assignment was performed using multiple computational approaches. For CRISPR spacer, tRNA and bacterial genome (BG) homology-based analyses, the RefSeq database of sequenced bacterial genomes was used (accessed 19 February 2018). For endemic bacterial population contig (BC) homology, total metagenomes were QCed, processed and assembled as mentioned earlier and only bacterial-assigned contigs ≥ 2.5 kb were kept for downstream analysis (RefSeq-based assignment, accessed 11 January 2017). For (A) CRISPR spacer homology: CRISPR spacers were extracted from RefSeq bacterial genomes using MinCED v0.2.0 (Skennerton, 2013) and queried against VCs using BLASTn v2.2.31+ (task = blastn-short, qcov_hsp_perc = 100, 2 mismatches/gaps allowed) (Altschul et al., 1990); (B) tRNA homology: tRNAs were extracted from RefSeq bacterial genomes using Aragorn v1.2.36 (-t) (Laslett and Canback, 2004) and queried against VCs using BLASTn (qcov_hsp_perc = 90, perc_identity = 90); (C) BG homology & (D) BC homology: VCs were queried against

BGs/BCs using BLASTn (perc_identity = 80, hits ≥ 1,000 nucleotides considered). BLASTn parameters used were based on parameters by Arkhipova et al. (2018), Paez-Espino et al. (2016) and Coutinho et al. (2017). Only the best BLAST hit was considered and collapsing of multiple CRISPR spacer, tRNA and BG hits per viral-bacterial assignment was performed. Taxonomic classification of hits obtained by (A), (B) and (C) was retrieved from NCBI using the taxonomizr v0.2.2 R package (Sherrill-Mix, 2017). Results of (D) were used to quantify and describe bacteriophage host population structures within sampled groundwater communities. Results of (A) and (C) were used to describe specific viral-bacterial interactions and investigate the occurrence of viral generalists in sequenced viromes. Viral-bacterial interactions were visualized using Cytoscape (Shannon et al., 2003). Representative viral generalists across samples were identified using cd-hit v4.6 (sequence identity threshold = 0.98, word_length = 11) (Li and Godzik, 2006). Only one duplicate contig was removed. The putative circularization of contigs of viral generalists was evaluated using VICA (Crits-Christoph, 2015). Closest relatives to viral generalists found were queried by aligning the contig subset against the viral nucleotide 'NR' database (taxid:10239) with the NCBI BLASTn tool (Johnson et al., 2008), using standard parameters and including regions of low complexity. Quantification of viral generalist abundance and occurrence on multiple samples was evaluated by aligning merged reads against contigs with BBMap v36.20 (% nucleotide identity = 0.99, random best mapping site selected) (Bushnell, 2016).

3. Results and Discussion

In total, 14 total metagenomes, 14 viromes and 14 16S rRNA amplicon datasets were generated using Illumina next generation sequencing technologies. Sequencing data corresponds to sampling events of contaminated groundwater ('C1 May 2016', 'C1 Aug 2016', 'C1 Nov 2016', 'C1 Feb 2017', 'C1 May 2017', 'C2 Mar 2017', 'C3 Apr 2017') and non-contaminated groundwater ('NC1 May 2016', 'NC1 Aug 2016', 'NC1 Nov 2016', 'NC1 Feb 2017', 'NC1 May 2017', 'NC2 Mar 2017', 'NC3 Apr 2017') from an old gasworks site. Results of the analysis of the next generation sequencing data and evaluation of community dynamics and viral-bacterial interactions follows below.

3.1 Groundwater Chemistry

Chemical data was collected at selected sampling stations by the local council in September/November 2015 and September 2016 (Supplementary Data A). In November 2015, the C1 sampling station showed the presence of polycyclic aromatic hydrocarbons (PAHs), benzene, toluene, ethylbenzene and xylene compounds (BTEX), and 1,2-dichloroethane (EDC) in its groundwater. Furthermore, C1 groundwater registered a pH of 9.52. The C1 sampling station was the closest to the predominant source of the contaminant plume, while C2 and C3 groundwater stations were located downstream and upstream of the majority of the plume, respectively. At the C2 sampling station high values of PAHs and BTEX were registered in September 2015 (*e.g.* 17,000 µg/mL total aromatic hydrocarbon compounds), however no EDC was found. Like C1, groundwater sampled at C2 also registered an alkaline pH (8.4). PAHs and BTEX were found in groundwater of the C3 station in September 2016. Here, concentrations were lower than those at C2 and lower/comparable to those at C1 (*e.g.* 390 µg/mL total aromatic hydrocarbon compounds), due to C3's upstream location in relation to the centre of the contaminant plume. The pH at C3 was registered at 6.85 and this value was closer to values registered for stations where no hydrocarbon groundwater contamination was found (6.96-7.23 for NC1, NC2 and NC3). Groundwater from the NC1 station was sampled twice by the local council and both in September 2015 and September 2016 no groundwater contamination was found. The pH at NC1 did not vary greatly (6.96 in September 2015 and 7.28 in September 2016). Other variations occurred however, such as changes in groundwater redox potential, levels of dissolved oxygen and concentration of sodium ions (Supplementary Data A).

3.2 Bacterial and Viral Community Diversities

3.2.1 Bacterial Communities

To study the bacterial diversity found at the gasworks site, amplicons of the 16S rRNA gene were generated and sequenced. A total of 1,107,323 amplicons with an average size ranging 475-504 bp per sample were obtained. Upon 16S amplicon data processing and OTU picking, 744,126 counts were assigned taxonomy and 23,573 OTUs were found. Amplicon counts ranged from 27,161 to 84,400 across samples and normalization by least sequencing depth was done. A total of 21297 OTUs were retained in the BIOM table (≈ 90%) and core diversity analysis was performed.

Principle coordinate analysis of OTUs showed that bacterial communities sampled from C1 and NC1 sampling stations clustered closely together while bacterial communities sampled from other stations across the site were placed further apart in the graph (Figure 2A). This suggested that groundwater bacterial community variation was greater across areas of the site than over time at the same sampling location. The C2 bacterial community was the one that most resembled the C1 sample group whilst NC3, NC2 and C3 bacterial communities most resembled those of the NC1 sample group. The variance was primarily explained by the x-axis (43%), likely relating to contaminant presence and pH variation. Only a small variance was observed in the y-axis (13%). C3, NC1, NC2 and NC3 samples had the most diverse bacterial communities (Shannon index H' = 6.62-7.21) when compared to C1 and C2 samples (H' = 3.84-5.22) (Supplementary Table B.1). Particularly, the bacterial community of C1 Aug 2016 had the lowest OTU richness (R = 2206) and evenness (E = 0.50) when compared to other C1 samples taken over the year (R = 2206-3174, E = 0.50-0.65), a deviation noticeable in Figure 2A. The lowest OTUs richness across the site was registered for the C2 Mar 2017 sample (R = 2130) despite its larger evenness (E = 0.63) when compared to C1 Aug 2016. The most diverse bacterial community was present at the NC2 Mar 2017 sampling station (R = 5596, E = 0.84).

Groundwater aquifers are not static and plumes of contamination may expand, migrate and mix (American Water Works Association, 2002). In light of this, and evidence that chemical changes have been actively occurring in groundwater at the site of study, we hypothesize that the local bacterial community at C1 was likely influenced by possible changes in groundwater chemistry or flow during August 2016. Over other time points, bacterial communities in C1 and NC1 were unlikely affected by any possible occurring variations in groundwater chemistry and/or migration of contaminant plumes.

The effect of pH as a critical influencer of microbial communities is well-established (Cho et al., 2016; Fierer and Jackson, 2006; Hartman et al., 2008; Lauber et al., 2009) and alkaline solutions were commonly used many decades ago in manufacturing gas plants (Thomas and Brinckerhoff, 2014). Foul lime, a rock solid material of high pH, is also commonly excavated from old gasworks sites (Thomas and Brinckerhoff, 2014). We hypothesize here that pH was likely to be the most important factor of bacterial community diversity in our site of study. This would explain why C3 bacterial communities were closer related to those of non-contaminated samples despite previous observations of PAH and BTEX contamination at this location. Furthermore, it would explain why the C1 sample group encompassed the most isolated group of samples in the PCoA (Figure 2A), with its closest bacterial community being that of C2 (registered pH of 9.52 and 8.4, respectively).

3.2.2 Viral Communities

To study the viral diversity found at the gasworks site, VLPs were isolated and viromes were sequenced. An excess of 51.6-150.8 M reads was generated per virome and 291,714-828,829 contigs were obtained per sample using metaSPAdes. A total of 84,974-719,249 ORFs per sample were predicted and annotated by homology to the Viral RefSeq protein database (9,401- 55,324 ORFs assigned). Virome annotations were normalized to 9,398 hits/sample and taxonomic assignment of virotypes found was performed.

Virotypes assigned by MEGAN's LCA algorithm (Huson et al., 2007) were used for viral diversity analysis (see Materials and Methods section 2.4 for cut-off values). Here, we found that viral diversity dissimilarities were consistent with bacterial diversity variations observed earlier (Figure 2B). Again here, the C2 viral community most resembled that of C1, and C3, whereas NC2 and NC3 most resembled viral communities of the NC1 sample group, with NC3 found to be highly similar to NC1 samples of May 2016, August 2016 and May 2017. NC1 samples from November 2016 were located further away from other NC1 samples on the y-axis of the graph (7.4%). Nonetheless, the majority of the variance was explained by the x-axis (67.4%). The dissimilarity of the viral community of C1 August 2016 from other C1 sample groups was visible along the x-axis. It has been previously shown that, next to temperature and nutrient availability, microbial diversity is the most important driver of viral abundance and production in ocean waters, as changes in the availability of hosts affects viruses that can survive in specific environments (Rowe et al., 2012). Overall, our results suggest that viral diversity found here mirrored the bacterial diversity found in groundwater, shaped by groundwater chemistry. Virotype diversity showed similar diversity metrics across samples (R = 689-813, E = 0.90-0.92, H' = 5.91-6.08) (Supplementary Table B.2). The highly similar evenness of virotypes at the site pointed out a low dominance of (previously-sequenced) viruses within the sampled microbial communities.

3.3 Bacterial and Viral Community Structures

3.3.1 Bacterial Communities

A total of 57 different phyla were found in sampled bacterial communities across the site. Unknown/unclassified bacterial amplicons represented 5-22% of counts across samples (Supplementary Figure B.1). In NC1, NC2, NC3 and C3 samples, Proteobacteria was the most abundant phyla throughout, representing 25-36% of assigned bacterial communities. Other abundant phyla at these stations included, for example, OD1 (2.2-9.1% in NC1), GN02 (6.5-21% in NC1, 13.7% in NC2, 11.2% in C3), Actinobacteria (5.3-11.3% in C1), Acidobacteria (18.2% in NC3), Chloroflexi (9.9% in NC3) and OP3 (14.6% in C3). In C1 groundwater communities, the most abundant phyla found was Bacteriodetes, representing up to 40% for majority of most sampled time points. In C1 Aug 2016 however, this was not the case. Instead, Proteobacteria represented 65.9% of the bacterial community. This was reflected in the dissimilarity of C1 Aug 2016 when compared to other C1 samples. Other abundant phyla at the C1 sampling station included Firmicutes (7.1-18.0%) and Chloroflexi (0.5-10.7%). The C2 bacterial community was most composed by a mix of Bacteriodetes (36.4%) and Proteobacteria (30.9%). This supported its location in the PCoA of Figure 2A.

Among the most abundant bacterial members at the site (Figure 3), a number have been linked to hydrocarbon biodegradation processes and/or hydrocarbon-contaminated environments. These include the Actinobacterial order iii1-15 (Morais et al., 2016), *Anaerolineaceae* (Kümmel et al., 2015; Liang et al., 2015; Rosenkranz et al., 2013), the Chloroflexi class GIF9 (Alfreider et al., 2002), the Elusimicrobiales order (Wright et al., 2017), *Thermoanaerobacteraceae* (Marozava et al., 2018), *Caulobacteraceae* (Martirani-Von Abercron et al., 2017; Morais et al., 2016; Yang et al., 2014, 2016), *Rhodospirillaceae* (Cui et al., 2008; Viñas et al., 2005), *Comamonadaceae* (Mattes et al., 2010; Morais et al., 2016; Yang et al., 2014), *Rhodocyclaceae* (Táncsics et al., 2018) and *Pseudomonadaceae* (Wald et al., 2015). The Actinobacterial order iii1-15 was particularly abundant in NC1 Feb 2017 (7.68%), NC1 May 17 (5.04%), NC2 (3.43%) and NC3 (12.81%) bacterial communities. *Rhodospirillaceae* was most abundant in NC1 (3.47-22.54%), NC2 (3.30%) and in NC3 (10.35%) bacterial communities. *Anaerolineaceae* and *Thermoanaerobacteraceae* families were most abundant in C1 (6.36-12.54% and 2.25-4.51%, respectively), and *Caulobacteraceae* and *Comamonadaceae* were most abundant in C2 (7.51% and 9.30%, respectively).

Tight ecological niches may oxidize organic pollutants to carbon dioxide by conducting aerobic respiration, denitrification and sulfate reduction at contaminant plume fringes, or by conducting iron and manganese reduction, and methanogenesis at the plume core (Meckenstock et al., 2015). Amongst the most abundant bacterial families at the site, some were associated with both hydrocarbon degradation and aforementioned processes. *Anaerolineaceae* has been described associated with methanogenesis and sulfate-reduction (Kümmel et al., 2015; Liang et al., 2015), *Thermoanaerobacteraceae* and *Caulobacteraceae* have been associated with sulfate reduction (Bagi et al., 2017), and, recently, *Comamonadaceae* has been implicated in a new mechanism of sulfur-driven iron reduction coupled to ammonium oxidation (Bao and Li, 2017).

Desulfobulbaceae members have been well-characterized as a sulfate-reducers (Mckew et al., 2013; Müller et al., 2009) and this family was abundant in C3 and NC1 May 2016 bacterial communities. *Geobacteraceae*, a family with sulfur and iron-reducing members (Caccavo et al., 1994; Lin et al., 2005), was also found abundant at the site (3.32% abundance in C3). Both *Geobacteraeae* and *Desulfobulbaceae* bacteria are able to perform long distance extracellular electron transport (Müller et al., 2016; Reguera et al., 2016). The abundance of *Desulfobulbaceae* and *Geobacteraceae* at the C3 sampling station could indicate an enhanced biodegradation capacity next to the putative plume fringe, based on sulfur cycling and long distance extracellular electron transport.

Sulfur oxidizers were abundant at the site. These include *Halothiobacillaceae* (Táncsics et al., 2018), *Hydrogenophilaceae* (Táncsics et al., 2018), *Rhodocyclaceae* (Táncsics et al., 2018), *Helicobacteraceae* (Ihara et al., 2017) and *Spirochaetaceae* (Zhang et al., 2017). *Halothiobacillaceae* was abundant at the C1 sampling station (0.56-4.21%) and *Spirochaetaceae* was most abundant in C3 bacterial communities. *Hydrogenophilaceae* and *Helicobacteraceae* families were highly abundant within the C1 bacterial communities over August 2016, representing 12.34% and 46.11% of the total bacterial community structure. This suggested that the decrease of bacterial diversity at C1 during August 2016 was due to an enrichment of two families involved in sulphur oxidation.

Abundant members found within sampled bacterial communities that have been linked to methanogenesis include, the actinobacterial order OPB41 (Robbins et al., 2016), the Methylophilales order (Redmond et al., 2010), *Porphyromonadaceae*

(Wang et al., 2016), *Hyphomicrobiaceae* (Beck et al., 2013; Karwautz et al., 2018; Osaka et al., 2008) and *Syntrophaceae* (Gray et al., 2011). *Porphyromonadaceae* was found particularly abundant in C1 (4.29-15.49%) and C2 communities (23.44%) and *Syntrophaceae* was found particularly abundant in C3 (7.71%) and NC2 (3.86%) communities.

A number of members of uncultured phyla were found abundant in sampled bacterial communities (Figure 3). These include members of candidate phyla GN02, OD1, OP3, OP11, TM6 and TM7. For example, the GKS2-174 class of GN02 was found highly abundant in C3 (10.83%), NC1 (4.36-19.48%), NC2 (13.14%) and NC3 sample communities (3.89%), and the TM7-3 class was most abundant in NC1 sample group (1.08-6.48%). Overall, 'NC' and C3 bacterial communities either presented similar or larger values for members of these phyla when compared to C1 and C2 communities. Some of these members have been associated with microbial denitrification, particularly OD1 classes ABY1 and ZB2, GN02, and koll11 class of OP3 (Hiller et al., 2015). *Nitrospiraceae*, a family of nitrite-oxidizers (Koch et al., 2015) was also found abundant in samples collected at the site. ML635J-40, an uncharacterized family previously found in extreme alkaline conditions, was found particularly abundant at the C1 sampling station (1.82-5.96%).

Sulfate and ammonia are known wastes of the gasworks production processes (Thomas and Brinckerhoff, 2014). The abundance of bacteria associated with sulfur and nitrogen metabolism at the site could be a result of this. The presence of not only sulfate-reducers but also methanogens in sampled groundwater communities, is further supported by redox values registered for groundwater at the site (-318 – 89 mV) (Supplementary Data A). The presence of sulfate-reducers, methanogens and several degraders at multiple sampling stations across site proposes that (A) bacteria found at the site were well-adjusted to environmental changes and that (B) the occurrence of dynamic groundwater flows and/or previous natural attenuation processes could be occurring over the decades.

With the availability of total metagenomics data, the presence of Archaea and Eukarya was inferred via SSU rRNA analysis and protein analysis. Archaeal members represented only up to 8.82% of total microbial communities at the site and most were methanogenic members of Euryarchaeota (data not shown). Lower eukaryotes represented only up to 0.64% of all microbes across samples (data not shown).

3.3.2 Viral Communities

Across the site, and over the yearlong sampling period, taxonomic assignments of viral communities were most represented by the Caudovirales bacteriophage families: *Siphoviridae* was the most abundant viral family in groundwater at the site (31-38%), followed by *Myoviridae* (16-20%), and *Podoviridae* (9-17%) (Supplementary Figure B.2). Research previously published by our group in a eutrophic freshwater lake in Northern Ireland showed *Podoviridae* populations as high as *Siphoviridae* (34.3% and 32.8%, respectively) (Skvortsov et al., 2016). In groundwater viral communities sampled here, distinct distributions were observed instead.

A total of 28-36% of *Siphoviridae,* 13-16% of *Myoviridae* and 8-15% of *Podoviridae* protein sequences were attributed to viruses yet to be classified (Supplementary Table B.3). Other unassigned and unclassified members of the Caudovirales order represented 13-15% and 1-2% of sequences, respectively (Supplementary Figure B.2). For remaining sequences, 3% were assigned to unclassified dsDNA phages,

≈ 1% to unclassified dsDNA viruses, 1-2% to unclassified bacterial viruses, and 4-5% remained unassigned at viral level. A small portion aligned to *Mimiviridae* (0.43-0.94%), *Phycodnaviridae* (1.16-2.49%) and others (0.42-0.76%). Some hits against ssDNA viruses (≤ 0.01%) were observed, despite exclusion of ssDNA viruses during metagenomic library preparation.

The diversity analysis of viruses with genera assigned revealed *T4virus* and *Lambdavirus* to be highly abundant across all samples (1.10-1.92% and 1.14-1.80%, respectively) (Figure 4A). *T4virus* were particularly abundant in C3 and 'NC' samples while *Lambdavirus* was particularly abundant in C1 and C2 samples. *Pamxvirus* were very abundant in C2 (1.68%) and *Chlorovirus,* predators of microscopic algae, were particularly enriched in NC3 (1.51%). *Bcep22virus* were most abundant in NC2 (1.00%) and NC1 samples (1.23-1.68%), and *Bpp1virus* widely abundant in 'NC' samples (1.16-1.68%), C2 (1.05%) and C3 (1.57%). *Bpp1virus* was also abundant in C1 during August 2016 (1.49% assigned sequences *vs.* 0.55-0.75% in other time points collected). Other genera that like *Bpp1virus* could explain C1 viral community dissimilarity during August 2016 include, for example, *Pamx74virus* (0.65% vs. 0.23-0.31%), *Slashvirus* (0.71% vs. 0.87-1.01%) and *Yuavirus* (0.98% vs. 0.48-0.57%). Similarly, genera variation that could explain the dissimilarity observed in the local NC1 community during November 2016 (Figure 2B) were, for example, *D3virus* (0.25% vs. 0.45-0.69%), *M12virus* (0.23% vs. 0.52-0.59%), *Prtbvirus* (0.61% vs. 0.82-0.91%) and *Xp10virus* (0.36% vs 0.72-0.85%).

Virotype dominance within local groundwater viral communities was investigated (Figure 4B). *Pelagiphages* have been described as the most abundant type of viruses across oceans and even the biosphere (Zhao et al., 2013). *Pelagiphages* were highly abundant in groundwater from this study, especially in 'NC' (1.96-2.26%), C2 (1.93%), C3 (2.24%), NC2 (2.00%) and NC3 (2.36%) viral communities. In groundwater form the C1 sampling station, *Pelagiphages* weren't as abundant however (0.92%-1.25%). Four *Pelagiphages* virotypes were found. Particularly, the *Pelagibacter* phage HTVC010P was highly abundant in 'NC', C2 and C3 communities, (1.43-1.67%). The *Pelagibacter* phage HTVC010P represented 0.74-0.90% of virotypes found over the year in C1.

Abundant virotypes found in groundwater samples from the C1 sampling station include the *Rhizobium* phage 16-3 (0.88-1.65%), *Bacillus* virus G (1.28-1.73%), *Bordetella* virus BBP1 (0.50-1.25%), *Cellulophaga* phage phi14:2 (0.58-1.32%), *Thermoanaerobacterium* phage THSA-485A (1.01-1.37%), *Paenibacillus* phage PG1 (0.96-1.74%) and *Geobacillus* virus E3 (0.83-1.29%). The increase of *Bordetella* virus BBP1 and *Rhizobium* phage 16-3 virotypes during Aug 2016 could also help explain its dissimilarity to C1 communities, along with the genera afore mentioned. The decrease of *Cellulophaga* phage phi14:2, *Paenibacillus* phage PG1, *Geobacillus* virus E3 virotypes during this time of the year could also be responsible for this. In NC1 sample groups, prominent virotypes observed included the *Bordetella* virus BBP1(0.93-1.37%), Myxococcus phage Mx8 (1.75-1.92%), *Rhodoferax* phage P26218 (2.02-2.54%), *Azospirillum* phage Cd (0.47-1.21%), *Caulobacter* phage CcrColossus (1.28-1.75%), *Rhizobium* phage 16-3 (0.91-2.04%), *Sinorhizobium* phage phiLM21 (0.69-1.38%) and *Synechococcus* phage S-CBS3 (1.01-1.28%). Here, the marked decrease of *Azospirillum* phage Cd, *Sinorhizobium* phage phiLM21 and *Rhizobium* phage 16-3 virotypes during November 2016 could contribute to the dissimilarity of this population when compared to other NC1 communities sampled over the year. Most of the virotypes found in high abundance in C1 and NC1 were also present in high abundance in C2, C3, NC2 and NC3 viral communities (Figure 4B). C2, however, additionally revealed a high abundance of the *Ralstonia* phage RSK1 (1.32%). Examples of other abundant virotypes found across the site include the *Vibrio* phage VvAW1, *Pseudomonas* phage AF and *Xanthomonas citri* phage

CP2 (0.38%-0.92% and 0.36%-0.80%, respectively). The Lough Neagh virome sequenced by our group (Skvortsov et al., 2016) revealed the high abundance of not only the *Pelagibacter* phage HTVC010P, but also the *Bordetella* virus BBP1, Myxococcus phage Mx8, *Rhizobium* phage 16-3 and *Vibrio* phage VvAW1 virotypes found here (Skvortsov et al., 2016). This present study sheds light into the abundance of these five virotypes not only in above ground freshwater but also in groundwater microbial communities.

Rhodoferax, *Rhizobium*, *Caulobacter, Ralstonia, Pseudomonas, Xanthomonas* and *Thermoanaerobacterium* bacterial species have been associated with the biodegradation of aromatic hydrocarbons (Aburto and Peimbert, 2011; Chatterjee and Bourquin, 1987; Latha and Mahadevan, 1997; Manickam et al., 2018; Marozava et al., 2018; Ryan et al., 2007; Wald et al., 2015) and the degrader families *Comamonadaceae* (*Rhodoferax*), *Thermoanaerobacteraceae (Thermoanaerobacterium), Caulobacteraceae (Caulobacter), Rhodospirillaceae (Azospirillum)* and *Pseudomonodaceae (Ralstonia, Pseudomonas)* were abundant at the site employed in this of study (see above). The abundance of virotypes infecting bacteria of these families suggests the possible on-site occurrence of bacteriophages with putative sways on natural attenuation processes and biodegradation strategies by disturbing the diversity and abundance of these defined bacterial degrader host populations.

3.4 Viral-Bacterial Associations

A range of 3120-10288 viral contigs (VCs) from sequenced viromes were used to identify bacteriophage host populations at the site of study. Four different computational methodologies were used. Using CRISPR Spacer homology, 17-42 (median x̃ = 23) VCs were assigned hosts and, similarly, using tRNA homology 8-26 (x̃ = 12) VCs were assigned hosts; using whole-contig homology against BGs from the RefSeq database, 1-21 (x̃ = 3) VCs had hosts assigned (Supplementary Table B.4). Because the RefSeq database is biased towards cultured organisms and because microbial communities from groundwater ecosystems have been marginally explored (Griebler and Lueders, 2009), total metagenomes from the site were sequenced and whole-contig homology against BCs was performed. Using this method, we were able to assigne hosts to 296-1948 VCs (x̃ = 1627) across datasets, finding putative hosts for 5.47-52.58% (x̃ = 27.0%) of VCs across samples. Other techniques assigned hosts for only 0.01-1.2% of VCs (x̃ =0.25%). Hence, BC homology data was used for description of broad host population structure dynamics at the site, and CRISPR Spacer homology and BG homology data was used for description of low level interactions and identification of viral generalists at the site, due to their higher fidelity of host species assignment (Edwards et al., 2016).

3.4.1 Host Community Structures

Inference of the host population structure at the site revealed Actinobacteria, Bacilli, Bacteroidia, Clostridia, Planctomycetia, Flavobacteriia and Proteobacteria classes as the most abundant for hosts of temperate phages (BC homology) (Figure 5A). For C1 samples, Clostridia was the most abundant host class found for VCs (23.03-26.67%), followed by Bacilli (12.76-13.22%), Bacteroidia (5.71-9.19%) and Deltaproteobacteria (8.63-10.53%). During August 2016, C1 prophage host populations were noticeably underrepresented by Bacteroidia (5.71% vs. 8.63-9.19% in other time points). Instead, Betaproteobacteria hosts were more abundant (6.59% vs. 2.90-4.15% in other time points). C2 host populations were likewise best represented by Clostridia (34.85%), Bacilli (12.76%), Bacteroidia (7.52%), Betaproteobacteria (7.74%) and Deltaproteobacteria (7.74%) members. NC1 and NC3 host populations were most abundant in Alphaproteobacteria (23.22-25.44%; 24.44%), Betaproteobacteria (13.22-15.38%; 20%), Gammaproteobacteria (11.99-16.22%; 23.70%) and Actinobacteria (9.49-10.57%; 24.44%). C3 and NC2 host populations were not only well represented by Betaproteobacteria (15.71%; 15.17%), Gammaproteobacteria (11.18%; 11.24%) and Alphaproteobacteria (15.11%; 19.10%) members, but also by Deltaproteobacteria members (19.10%; 23.56%). Other classes, such as Planctomycetia and Flavobacteriia were also somewhat abundant amongst lysogenic bacteria host populations (0.46-5.95% and 0.91-3.30%, respectively) despite families of these classes not being amongst the most abundant at the site (Figure 2). Planctomycetia represent a class of bacteria commonly found in freshwater (Fuerst and Sagulenko, 2011) and Flavobacteriia members have been associated with the degradation of PAHs (Juhasz and Naidu, 2000; Kappell et al., 2014; Trzesicka-Mlynarz and Ward, 1995). By targeting Proteobacteria and Flavobacteriia members, bacteriophages could impact biodegradation rates at the site during cell lysis and viral particle release. Other classes of degraders found amongst putative prophage hosts at the site include, for example, the Anaerolineae class, although a relatively low VC assignment was observed (0.37%-1.77%). Overall, shifts in prophage host populations described here are explained by the dissimilarities observed in bacterial and viral communities reported earlier (Figure 2).

The dynamics of putative bacteriophage-host interactions at the site was investigated (Figure 5B). Host sequences (BCs) assigned to viruses of the C1 sample group were most found within C1 microbial communities themselves (13.29-25.05%), totaling 95.84-96.94% of all matches. Matches to other communities represented only 3.46-4.97% of all assignments for C1. In the NC1 sample group, most host sequences were also within the same microbial communities (9.29-42.74%), totaling 96.96-98.06% of assignments. BCs of other communities accounted only for 1.95-2.92% of NC1 BC assignments. This suggests the occurrence of well-defined ecological niches at the site. Particularly, the C1 location represents a well-defined ecological niche near the centre of the contaminant plume.

Similar to C1 and NC1, most of the host sequences identified for the C3 and NC2 viral communities were found at C3 and NC2 sampling stations (72.98% and 69.78%, respectively). This indicated that upstream the plume centre (C3 location) a somewhat defined ecological niche, distinct from C1, was also found. Downstream the plume centre (C2 location), however, host sequences found originated not only within-community (22.33%) but also from the C1 sampling station (9.5-12.52% across C1 time points). The same was true for the viral communities of NC3, where only 35.85% of assigned BCs were found in NC3's own microbial community. Here, up to 11.15% of NC3 hits were found at C3, NC2, and across NC1 samples. These results could be a reflection of dynamic groundwater flow and/or dynamic groundwater mixing at the site, where some bacteriophages may be found across locations but bacterial hosts may not be able to adapt and prosper in new environmental conditions. The evidence for possible dynamic groundwater flows at the site of study could further justify the variance observed at C1 during August 2016, particularly if changes to the water table occurred.

3.4.2 Broad Host Range Interactions

Host-bacteriophage assignments were discriminated at bacterial species level, and interactions between VCs and putative host species across the site were projected (Figure 6). Thirty-six unique viral generalists, *i.e.* viruses infecting more than one bacterial species, were found and their hosts were described (Table 1). Seventeen generalists were described by CRISPR spacer homology, 17 by BG homology and two by both methods. Seventeen generalists were classified as multi-species generalists and nineteen were classified as multi-genera generalists due to putatively infecting species from different genera (and above). Contig size ranged from 2531 bp to 78895 bp (x̃ = 5541) (Supplementary Table B.5). When possible, generalists were classified as *Podoviridae* (one/36), *Myoviridae* (two/36), *Siphoviridae* (8/36) and Caudovirales (11/36) members using CAT (Cambuy et al., 2016).

A total of 11 generalists described here were found to putatively infect members of the *Pseudomonas* genus (Table 1). Particularly, BGW-G9 aligned to RefSeq genomes ranging 23 taxonomic assignments, most of which were represented by pseudomonads. These included strains of three *Pseudomonas* species *(Pseudomonas aeruginosa, Pseudomonas denitrificans* and *Pseudomonas pseudoalcaligenes)* and 17 unclassified *Pseudomonas* isolates. Other putative hosts of BGW-G9 were members of the *Polycyclovorans algicola, Methylocaldum szegediense,* and *Candidatus Magnetobacterium casensis* species. Overall, putative host species for BGW-G9 were represented by an excess of 1394 CRISPR Spacer and 137 BC hits against the RefSeq database of sequenced bacterial genomes. The closest relative found for BGW-G9 in the NR database was the *Pseudomonas* phage JBD26 (88% query cover, 98% identity) (Supplementary Table B.6).

BGW-G23 and BGW-G32 generalists were also represented by a large array of hosts ranging 20 and 23 different taxonomic assignments across the *Acinetobacter* genus. Contigs of both BGW-G23 and BGW-G32 represented complete circular phage genomes (Supplementary Table B.5). BGW-G23 was particularly abundant in groundwater at the C2 sampling station, with 21,413 counts per million reads assigned. Other generalists at the site were only represented by up to 1,086 counts per million reads in groundwater across the site (Supplementary Figure B.3).

The occurrence of putative generalists such as BGW-G9, BGW-G23 and BGW-G32 could have a marked impact in natural attenuation processes and implementation of bioremediation strategies at the site of study, as they are putatively able to singularly infect and modulate populations of several species with biodegradative capacity, *i.e. Pseudomonas* sp., *Polycyclovorans algicola* and *Acinetobacter* sp. (Gutierrez et al., 2013; Simarro et al., 2013; Wald et al., 2015).

Bacteria, with (strain-level) relatives able of biodegradation, that could also be affected by viral generalists found here include, for example, *Thermoanaerobacter* spp. (by BGW-G1 and BGW-20), *Porphyromonadaceae* spp. (by BGW-29), *Burkholderia* spp. (by BGW-8, BGW-11 and BGW-35), *Mycobacterium* spp. (by BGW-16), a *Xanthomonas* sp. (by BGW-35), a *Comamonadaceae* sp. (by BGW-10), a *Flavobacterium* sp. (by BWG-G5 and BWG-G7), a *Raoultella* sp. (by BGW-8), a *Caulobacter* sp. (by BGW-15), and a *Hydrocarboniphaga* sp. (by BGW-34) (Burback and Perry, 1993; Chatterjee and Bourquin, 1987; Manickam et al., 2007; Mattes et al., 2010; Palleroni et al., 2004; Ping et al., 2017; Poi et al., 2018; Revathy et al., 2015; Simarro et al., 2013). *Thermoanaerobacteraceae, Comamonadaceae, Porphyromonadaceae, Caulobacteraceae and Pseudomonadaceae* members were particularly abundant at the site of study (Figure 3) and their putative natural attenuation processes could be particularly impacted by some of the viral generalists described here. Furthermore, while putatively contributing to a wider decline in degrader's biomass, viral generalists found here could also have a wider role in the viral shunt of microbial communities (Weinbauer and Rassoulzadegan, 2004).

4. Conclusion

By conducting a yearlong metagenomic study on viruses and bacteria of groundwater from an old gasworks site, we were able to observe that community changes were greater across the site than over time at the same sampling station. We hypothesize that this could be due to the known differences in pH, and to a lesser degree, contaminants at the site. Non-surprisingly, we observed that viral communities at the site mirrored the diversity of the bacterial communities sampled. Hydrocarbon degraders were abundant within sampled microbial communities and virotypes of predators of bacterial degraders were also found. By further studying viral-bacterial interactions occurring at site we were able to pinpoint host populations and also describe where discrete host-phage interactions were taking place. A number of viral generalists with putative impact in biodegradation processes were also found. Overall, findings reported here support the employment of phage research during the development of bioremediation strategies.

In this study, we shed a new light not only on the putative impact of local bacteriophage communities in natural attenuation and bioremediation processes but also onto the viral community structures of an environment not addressed before.

Appendix A. Supplementary data A

Chemical description of groundwater at the site of study (.xlsx).

Appendix B. Supplementary data B

Support tables and figures for bacterial and viral analyses presented (.docx).

Acknowledgements

We would like to thank Brendan McLean for supporting access to sampling site and for sharing groundwater chemistry for the site of study. Ricardo Costeira, Rory Doherty, Christopher CR Allen, Michael J Larkin and Leonid A Kulakov were supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 643087 REMEDIATE (Improved decision-making in contaminated land site investigation and risk assessment).

References

Aburto A, Peimbert M. Degradation of a benzene-toluene mixture by hydrocarbon-adapted bacterial communities. Ann Microbiol 2011;61:553–62. doi:10.1007/s13213-010-0173-6.

Alfreider A, Vogt C, Babel W. Microbial diversity in an in situ reactor system treating monochlorobenzene contaminated groundwater as revealed by 16S ribosomal DNA analysis. Syst Appl Microbiol 2002;25:232–40. doi:10.1078/0723-2020-00111.

Altschul SF, Gish W, Miller W, Myers WE, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:402–10.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010; Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Aracic S, Manna S, Petrovski S, Wiltshire JL, Mann G, Franks AE. Innovative biological approaches for monitoring and improving water quality. Front Microbiol 2015;6:826. doi:10.3389/fmicb.2015.00826.

Arkhipova K, Skvortsov T, Quinn JP, McGrath JW, Allen CCR, Dutilh BE, et al. Temporal dynamics of uncultured viruses: A new dimension in viral diversity. ISME J 2018;12:199–211. doi:10.1038/ismej.2017.157.

American Water Works Association. Groundwater Management and Protection. Groundwater (M21), 3rd Ed, 2002, p. 33-41.

Bagi Z, Ács N, Böjti T, Kakuk B, Rákhely G, Strang O, et al. Biomethane: The energy storage, platform chemical and greenhouse gas mitigation target. Anaerobe 2017;46:13–22. doi:10.1016/j.anaerobe.2017.03.001.

Bao P, Li GX. Sulfur-Driven Iron Reduction Coupled to Anaerobic Ammonium Oxidation. Environ Sci Technol 2017;51:6691–8. doi:10.1021/acs.est.6b05971.

Beck DAC, Kalyuzhnaya MG, Malfatti S, Tringe SG, Glavina del Rio T, Ivanova N, et al. A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the Methylococcaceae and the Methylophilaceae. PeerJ 2013;1:e23. doi:10.7717/peerj.23.

Brandon E. The Nature and Extent of Site Contamination. Glob Approch Site Cont Law. 2013, p. 11-39.

Breitbart M. Marine Viruses: Truth or Dare. Ann Rev Mar Sci 2012;4:425–48. doi:10.1146/annurev-marine-120709-142805.

Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2014;12:59–60. doi:10.1038/nmeth.3176.

Burback BL, Perry JJ. Biodegradation and biotransformation of groundwater pollutant mixtures by Mycobacterium vaccae. Appl Environ Microbiol 1993;59:1025–9.

Bushnell B. BBMap short read aligner. 2016; Available online at: http://sourceforge.net/projects/bbmap

Cambuy DD, Coutinho FH, Dutilh BE. Contig annotation tool CAT robustly classifies assembled metagenomic contigs and long sequences. 2016. doi:10.1101/072868.

Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann ML, Brüssow H. Phage as agents of lateral gene transfer. Curr Opin Microbiol 2003;6:417–24. doi:10.1016/S1369-5274(03)00086-9.

Caporaso JG, Bittinger K, Bushman FD, Desantis TZ, Andersen GL, Knight R. PyNAST: A flexible tool for aligning sequences to a template alignment. Bioinformatics 2010;26:266–7. doi:10.1093/bioinformatics/btp636.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high- throughput community sequencing data. Nat Methods 2010;7:335–6. doi:10.1038/nmeth0510-335.

Caccavo F, Lonergan DJ, Lovley DR, Davis M. Acetate- Oxidizing Dissimilatory Metal-Reducing Microorganism. Microbiology 1994;60:3752–9.

Chatterjee DK, Bourquin AW. Metabolism of aromatic compounds of Caulobacter crescentus. J Bacteriol 1987;169:1993–6. doi:10.1128/jb.169.5.1993-1996.1987.

Cho S, Kim M, Lee Y. Effect of pH on soil bacterial diversity. J Ecol Environ 2016;40:10. doi:10.1186/s41610-016-0004-1.

Clokie MR, Millard AD, Letarov A V, Heaphy S. Phages in nature. Bacteriophage 2011;1:31–45. doi:10.4161/bact.1.1.14942.

Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. Nat Commun 2017;8:1–12. doi:10.1038/ncomms15955.

Crits-Christoph A. VIral and Circular content from metAgenomes (VICA). 2015; Available online at: https://github.com/alexcritschristoph/VICA

Cui Z, Lai Q, Dong C, Shao Z. Biodiversity of polycyclic aromatic hydrocarbon-degrading bacteria from deep sea sediments of the Middle Atlantic Ridge 2008;10:2138–49. doi:10.1111/j.1462-2920.2008.01637.x.

Danielopol D, Griebler C, Gunatilaka A, Notenboom J. Present state and future prospects for groundwater ecosystems. Environ Conserv 2003;30:1–27. doi:10.1017/S0376892903000.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol 2006;72:5069–72. doi:10.1128/AEM.03006-05.

Devarapalli P, Kumavath RN. Metagenomics — A Technological Drift in Bioremediation. Intech, 2015, p. 73–91.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26:2460–1. doi:10.1093/bioinformatics/btq461.

Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev 2016;40:258–72. doi:10.1093/femsre/fuv048.

Eydal HSC, Jägevall S, Hermansson M, Pedersen K. Bacteriophage lytic to Desulfovibrio aespoeensis isolated from deep groundwater. ISME J 2009;3:1139–47. doi:10.1038/ismej.2009.66.

Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. Proc Natl Acad Sci U S A 2006;103:626–31. doi:10.1073/pnas.0507535103.

Fuerst JA, Sagulenko E. Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. Nat Rev Microbiol 2011;9:403–13.

Gray ND, Sherry A, Grant RJ, Rowan AK, Hubert CRJ, Callbeck CM, et al. The quantitative significance of Syntrophaceae and syntrophic partnerships in methanogenic degradation of crude oil alkanes. Environ Microbiol 2011;13:2957–75. doi:10.1111/j.1462-2920.2011.02570.x.

Griebler C, Lueders T. Microbial biodiversity in groundwater ecosystems. Freshw Biol 2009;54:649–77. doi:10.1111/j.1365-2427.2008.02013.x.

Gutierrez T, Green DH, Nichols PD, Whitman WB, Semple KT, Aitken MD. Polycyclovorans algicola gen. nov., sp. nov., an aromatic-hydrocarbon- degrading marine bacterium found associated with laboratory cultures of marine phytoplankton. Appl Environ Microbiol 2013;79:205–14. doi:10.1128/AEM.02833-12.

Hartman WH, Richardson CJ, Vilgalys R, Bruland GL. Environmental and anthropogenic controls over bacterial communities in wetland soils. Proc Natl Acad Sci 2008;105:17842–7. doi:10.1073/pnas.0808254105.

Hiller KA, Foreman KH, Weisman D, Bowen JL. Alter Bacterial Community Composition and Aquifer Redox Conditions 2015;81:7114–24. doi:10.1128/AEM.01986-15.

Huson D, Auch A, Qi J, Schuster S. MEGAN analysis of metagenome data. Gennome Res 2007;17:377–86. doi:10.1101/gr.5969107.

Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput Biol 2016;12:1–12. doi:10.1371/journal.pcbi.1004957.

Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11. doi:10.1186/1471-2105-11-119.

Ihara H, Hori T, Aoyagi T, Takasaki M, Katayama HY. Sulfur-oxidizing bacteria mediate microbial community succession and element cycling in launched marine sediment. Front Microbiol 2017;8:1–11. doi:10.3389/fmicb.2017.00152.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. Nucleic Acids Res 2008;36:5–9. doi:10.1093/nar/gkn201.

Joshi N, Fass J. sickle - A windowed adaptive trimming tool for FASTQ files using quality. 2011; Available online at: https://github.com/najoshi/sickle

Juhasz AL, Naidu R. Bioremediation of high molecular weight polycyclic aromatic hydrocarbons: a review of the microbial degradation of benzo[a]pyrene. Int Biodeterior Biodegradation 2000;45:57–88. doi:10.1016/S0964-8305(00)00052-4.

Kappell AD, Wei Y, Newton RJ, van Nostrand JD, Zhou J, McLellan SL, et al. The polycyclic aromatic hydrocarbon degradation potential of Gulf of Mexico native coastal microbial communities after the Deepwater Horizon oil spill. Front Microbiol 2014;5:1–13. doi:10.3389/fmicb.2014.00205.

Karwautz C, Kus G, Stöckl M, Neu TR, Lueders T. Microbial megacities fueled by methane oxidation in a mineral spring cave. ISME J 2018;12:87–100. doi:10.1038/ismej.2017.146.

Koch H, Lücker S, Albertsen M, Kitzinger K, Herbold C, Spieck E, et al. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. Proc Natl Acad Sci 2015;112:11371–6. doi:10.1073/pnas.1506533112.

Kümmel S, Herbst FA, Bahr A, Duarte M, Pieper DH, Jehmlich N, et al. Anaerobic naphthalene degradation by sulfatereducing Desulfobacteraceae from various anoxic aquifers. FEMS Microbiol Ecol 2015;91:fiv006. doi:10.1093/femsec/fiv006.

Kyle JE, Eydal HSC, Ferris FG, Pedersen K. Viruses in granitic groundwater from 69 to 450 m depth of the Äspö hard rock laboratory, Sweden. ISME J 2008;2:571–4. doi:10.1038/ismej.2008.18.

Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 2004;32:11–6. doi:10.1093/nar/gkh152.

Latha S, Mahadevan A. Role of rhizobia in the degradation of aromatic substances. World J Microbiol Biotechnol 1997;13:601–7. doi:10.1023/A:1018598200187.

Lauber CL, Hamady M, Knight R, Fierer N. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 2009;75:5111–20. doi:10.1128/AEM.00335-09.

Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9. doi:10.1093/bioinformatics/btl158.

Liang B, Wang LY, Mbadinga SM, Liu JF, Yang SZ, Gu JD, et al. Anaerolineaceae and Methanosaeta turned to be the dominant microorganisms in alkanes-dependent methanogenic culture after long-term of incubation. AMB Express 2015;5:37. doi:10.1186/s13568-015-0117-4.

Lin B, Braster M, Van Breukelen BM, Van Verseveld HW, Westerhoff H V, Röling WFM. Geobacteraceae community composition is related to hydrochemistry and biodegradation in an iron-reducing aquifer polluted by a neighboring landfill. Appl Environ Microbiol 2005;71:5983–91. doi:10.1128/AEM.71.10.5983-5991.2005.

Majone M, Verdini R, Aulenta F, Rossetti S, Tandoi V, Kalogerakis N, et al. In situ groundwater and sediment bioremediation: Barriers and perspectives at European contaminated sites. N Biotechnol 2015;32:133–46. doi:10.1016/j.nbt.2014.02.011.

Manickam N, Misra R, Mayilraj S. A novel pathway for the biodegradation of γ-hexachlorocyclohexane by a Xanthomonas sp. strain ICH12. J Appl Microbiol 2007;102:1468–78. doi:10.1111/j.1365-2672.2006.03209.x.

Marie V, Lin J. Viruses in the environment - presence and diversity of bacteriophage and enteric virus populations in the Umhlangane River, Durban, South Africa. J Water Health 2017;15:966–81. doi:10.2166/wh.2017.066.

Marozava S, Mouttaki H, Müller H, Laban NA, Probst AJ, Meckenstock RU. Anaerobic degradation of 1-methylnaphthalene by a member of the Thermoanaerobacteraceae contained in an iron-reducing enrichment culture. Biodegradation 2018;29:23–39. doi:10.1007/s10532-017-9811-z.

Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJournal 2011;17:10. doi:10.14806/ej.17.1.200.

Martirani-Von Abercron SM, Marín P, Solsona-Ferraz M, Castañeda-Cataña MA, Marqués S. Naphthalene biodegradation under oxygen-limiting conditions: community dynamics and the relevance of biofilm-forming capacity. Microb Biotechnol 2017;10:1781–96. doi:10.1111/1751-7915.12842.

Mattes TE, Alexander AK, Coleman N V. Aerobic biodegradation of the chloroethenes: Pathways, enzymes, ecology, and evolution. FEMS Microbiol Rev 2010;34:445–75. doi:10.1111/j.1574-6976.2010.00210.x.

Mckew BA, Dumbrell AJ, Taylor JD, Mcgenity TJ, Underwood GJC. Differences between aerobic and anaerobic degradation of microphytobenthic biofilm-derived organic matter within intertidal sediments. FEMS Microbiol Ecol 2013;84:495–509. doi:10.1111/1574-6941.12077.

Meckenstock RU, Elsner M, Griebler C, Lueders T, Stumpp C, Aamand J, et al. Biodegradation: Updating the Concepts of Control for Microbial Cleanup in Contaminated Aquifers. Environ Sci Technol 2015;49:7073–81. doi:10.1021/acs.est.5b00715.

Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: Evaluation of metagenome assemblies. Bioinformatics 2016;32:1088–90. doi:10.1093/bioinformatics/btv697.

Morais D, Pylro V, Clark IM, Hirsch PR, Tótola MR. Responses of microbial community from tropical pristine coastal soil to crude oil contamination. PeerJ 2016;4:e1733. doi:10.7717/peerj.1733.

Müller H, Bosch J, Griebler C, Damgaard LR, Nielsen LP, Lueders T, et al. Long-distance electron transfer by cable bacteria in aquifer sediments. ISME J 2016;10:2010–9. doi:10.1038/ismej.2015.250.

Müller S, Vogt C, Laube M, Harms H, Kleinsteuber S. Community dynamics within a bacterial consortium during growth on toluene under sulfate-reducing conditions. FEMS Microbiol Ecol 2009;70:586–96.

doi:10.1111/j.1574-6941.2009.00768.x.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res 2017;27:824–34. doi:10.1101/gr.213959.116.

O'Brien E, Nakyazze J, Wu H, Kiwanuka N, Cunningham W, Kaneene JB, et al. Viral diversity and abundance in polluted waters in Kampala, Uganda. Water Res 2017;127:41–9. doi:10.1016/j.watres.2017.09.063.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Mcglinn D, Minchin PR, et al. vegan: Community Ecology Package. Avialable online at: https://CRAN.R-project.org/package=vegan 2018.

Osaka T, Ebie Y, Tsuneda S, Inamori Y. Identification of the bacterial community involved in methane-dependent denitrification in activated sludge using DNA stable-isotope probing. FEMS Microbiol Ecol 2008;64:494–506. doi:10.1111/j.1574-6941.2008.00473.x.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature 2016;536:425–30. doi:10.1038/nature19094.

Palleroni NJ, Port AM, Chang HK, Zylstra GJ. Hydrocarboniphaga effusa gen. nov., sp. nov., a novel member of the γ-Proteobacteria active in alkane and aromatic hydrocarbon degradation. Int J Syst Evol Microbiol 2004;54:1203–7. doi:10.1099/ijs.0.03016-0.

Pan D, Nolan J, Williams KH, Robbins MJ, Weber KA. Abundance and Distribution of Microbial Cells and Viruses in an Alluvial Aquifer. Front Microbiol 2017;8:1–11. doi:10.3389/fmicb.2017.01199.

Panagos P, Hiederer R, Van Liedekerke M, Bampa F. Contaminated Sites in Europe: Review of the Current Situation Based on Data Collected through a European Network. J Environ Public Health 2013;2013:158764. doi:DOI 10.1016/j.ecolind.2012.07.020.

Ping L, Guo Q, Chen X, Yuan X, Zhang C, Zhao H. Biodegradation of pyrene and benzo[a]pyrene in the liquid matrix and soil by a newly identified Raoultella planticola strain. 3 Biotech 2017;7:56. doi:10.1007/s13205-017-0704-y.

Poi G, Shahsavari E, Aburto-Medina A, Mok PC, Ball AS. Large scale treatment of total petroleum-hydrocarbon contaminated groundwater using bioaugmentation. J Environ Manage 2018;214:157–63. doi:10.1016/j.jenvman.2018.02.079.

Redmond MC, Valentine DL, Sessions AL. Identification of novel methane-, ethane-, and propane-oxidizing bacteria at marine hydrocarbon seeps by stable isotope probing. Appl Environ Microbiol 2010;76:6412–22. doi:10.1128/AEM.00271-10.

Reguera G, Nevin KP, Nicoll JS, Covalla SF, Woodard TL, Lovley DR. Biofilm and nanowire production leads to increased current in Geobacter sulfurreducens fuel cells. Appl Environ Microbiol 2006;72:7345–8. doi:10.1128/AEM.01444-06.

Revathy T, Jayasri MA, Suthindhiran K. Biodegradation of PAHs by Burkholderia sp. VITRSB1 Isolated from Marine Sediments. Scientifica (Cairo) 2015;2015:9. doi:10.1155/2015/867586.

Robbins SJ, Evans PN, Parks DH, Golding SD, Tyson GW. Genome-centric analysis of microbial populations enriched by hydraulic fracture fluid additives in a coal bed methane production well. Front Microbiol 2016;7:731. doi:10.3389/fmicb.2016.00731.

Rohwer F, Prangishvili D, Lindell D. Roles of viruses in the environment. Environ Microbiol 2009;11:2771–4. doi:10.1111/j.1462-2920.2009.02101.x.

Rosenkranz F, Cabrol L, Carballa M, Donoso-Bravo A, Cruz L, Ruiz-Filippi G, et al. Relationship between phenol degradation efficiency and microbial community structure in an anaerobic SBR. Water Res 2013;47:6739–49. doi:10.1016/j.watres.2013.09.004.

Rowe JM, Debruyn JM, Poorvin L, Lecleir GR, Johnson ZI, Zinser ER, et al. Viral and bacterial abundance and production in the Western Pacific Ocean and the relation to other oceanic realms. FEMS Microbiol Ecol 2012;79:359–70. doi:10.1111/j.1574-6941.2011.01223.x.

Ryan MP, Pembroke JT, Adley CC. Ralstonia pickettii in environmental biotechnology: Potential and applications.

J Appl Microbiol 2007;103:754–64. doi:10.1111/j.1365-2672.2007.03361.x.

Shannon P, Markiel A, Owen Ozier 2, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003:2498–504. doi:10.1101/gr.1239303.metabolite.

Sherrill-Mix S. taxonomizr: Functions to Work with NCBI Accessions and Functions. 2017; Available online at: https://CRAN.R-project.org/package=taxonomizr

Simarro R, González N, Bautista LF, Molina MC. Biodegradation of high-molecular-weight polycyclic aromatic hydrocarbons by a wood-degrading consortium at low temperatures. FEMS Microbiol Ecol 2013;83:438–49. doi:10.1111/1574-6941.12006.

Skennerton CT. MinCED - Mining CRISPRs in Environmental Datasets. 2013; Available online at: https://github.com/ctSkennerton/minced

Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, et al. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. PLoS One 2016;11:e0150361. doi:10.1371/journal.pone.0150361.

Smith RJ, Jeffries TC, Roudnew B, Seymour JR, Fitch AJ, Simons KL, et al. Confined aquifers as viral reservoirs. Environ Microbiol Rep 2013;5:725–30. doi:10.1111/1758-2229.12072.

Soergel DAW, Dey N, Knight R, Brenner SE. Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. ISME J 2012;6:1440–4. doi:10.1038/ismej.2011.208.

Táncsics A, Szalay A, Farkas M, Benedek T, Szoboszlay S, Szabó I, et al. Stable isotope probing of hypoxic toluene degradation at the Siklo's aquifer reveals prominent role of Rhodocyclaceae. FEMS Microbiol Ecol 2018;94:fiy088. doi:10.1093/femsec/fiy088.

Thomas DR, Brinckerhoff P. Gasworks Profile A: The History and Operation of Gasworks (Manufactured Gas Plants) in Britain, 2014.

Thurber R V, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc 2009;4:470–83. doi:10.1038/nprot.2009.10.

Thurber RV. Methods in viral metagenomics. Handb. Mol. Microb. Ecol. II Metagenomics Differ. Habitats, 2011, p. 15–24.

Trzesicka-Mlynarz D, Ward OP. Degradation of polycyclic aromatic hydrocarbons (PAHs) by a mixed culture and its component pure cultures, obtained from PAH-contaminated soil. Can J Microbiol 1995;41:470–6. doi:10.1139/m95-063.

Viñas M, Sabaté J, Espuny MJ, Anna M, Vin M. Bacterial Community Dynamics and Polycyclic Aromatic Hydrocarbon Degradation during Bioremediation of Heavily Creosote-Contaminated Soil Bacterial Community Dynamics and Polycyclic Aromatic Hydrocarbon Degradation during Bioremediation of Heavily Creosote. Appl Environ Microbiol 2005, 2005;71:7008–18. doi:10.1128/AEM.71.11.7008.

Wald J, Hroudova M, Jansa J, Vrchotova B, Macek T, Uhlik O. Pseudomonads rule degradation of polyaromatic hydrocarbons in aerated sediment. Front Microbiol 2015;6:1268. doi:10.3389/fmicb.2015.01268.

Wang Q, Liang Y, Zhao P, Li QX, Guo S, Chen C. Potential and optimization of two-phase anaerobic digestion of oil refinery waste activated sludge and microbial community study. Sci Rep 2016;6:1–10. doi:10.1038/srep38245.

Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? Environ Microbiol 2004;6:1–11. doi:10.1046/j.1462-2920.2003.00539.x.

Wommack KE, Colwell RR. Virioplankton: Viruses in Aquatic Ecosystems. Microbiol Mol Biol Rev 2000;64:69–114. doi:10.1128/MMBR.64.1.69-114.2000.

Wright J, Kirchner V, Bernard W, Ulrich N, McLimans C, Campa MF, et al. Bacterial community dynamics in dichloromethane-contaminated groundwater undergoing natural attenuation. Front Microbiol 2017;8:2300. doi:10.3389/fmicb.2017.02300.

Yang S, Wen X, Shi Y, Liebner S, Jin H, Perfumo A. Hydrocarbon degraders establish at the costs of microbial richness, abundance and keystone taxa after crude oil contamination in permafrost environments. Sci Rep 2016;6:37473. doi:10.1038/srep37473.

Yang S, Wen X, Zhao L, Shi Y, Jin H. Crude oil treatment leads to shift of bacterial communities in soils from the deep active layer and upper permafrost along the China-Russia Crude Oil Pipeline route. PLoS One 2014;9:e96552. doi:10.1371/journal.pone.0096552.

Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014;30:614–20. doi:10.1093/bioinformatics/btt593.

Zhang Y, Wang X, Zhen Y, Mi T, He H, Yu Z. Microbial diversity and community structure of sulfate-reducing and sulfur-oxidizing bacteria in sediment cores from the East China Sea. Front Microbiol 2017;8:2133. doi:10.3389/fmicb.2017.02133.

Zhao Y, Temperton B, Thrash JC, Schwalbach MS, Vergin KL, Landry ZC, et al. Abundant SAR11 viruses in the ocean. Nature 2013;494:357–60. doi:10.1038/nature11921.

Table 1. Description of host species found for viral generalists present in sequenced viromes. Host assignment based on CRISPR Spacer homology (CRISPR) and BG hits to the RefSeq bacterial genomes database. 'MS' indicates multi-species generalists and 'MG' indicates multi-genera generalists (and above).

ID	Group	Method	Hits	Taxa	Status	Putative Host Species
BGW-G1	Viruses	CRISPR	3	3	MS	*Thermoanaerobacter* sp. X514, X513, X561
BGW-G2	Caudovirales	CRISPR	2	2	MG	*Halotalea alkalilenta; Halomonas* sp. 141
BGW-G3	Viruses	CRISPR	2	2	MG	*Prosthecobacter debontii; Rubellimicrobium thermophilum*
BGW-G4	Caudovirales	BG	2	2	MS	*Desulfotomaculum gibsoniae; Desulfotomaculum arcticum*
BGW-G5	Viruses	BG	2	2	MG	*Flavobacterium cyanobacteriorum; Hydrotalea flava*
BGW-G6	Viruses	BG	2	2	MS	*Saccharomonospora cyanea; Saccharomonospora* sp. LRS4.154
BGW-G7	Viruses	BG	3	3	MG	*Flavobacterium cyanobacteriorum; Hydrotalea flava; Chryseobacterium* sp. RU37D
BGW-G8	*Myoviridae*	CRISPR	10	5	MG	*Pseudomonas aeruginosa; Salmonella enterica; Burkholderia gladioli; Raoultella planticola; Halomonas* sp. 141
BGW-G9	*Siphoviridae*	CRISPR / BG	1394 / 137	23	MG	*Pseudomonas aeruginosa; Pseudomonas denitrificans; Pseudomonas pseudoalcaligenes; Pseudomonas* sp. P179, ADP, EGD-AKN5, HMSC072F09, HMSC064G05, HMSC065H02, HMSC063H08, HMSC073F05, HMSC065H01, HMSC057H01, HMSC058B07, HMSC059F05, HMSC060G02, HMSC061A10, HMSC070B12, HMSC058C05, HMSC11A05*; Polycyclovorans algicola; Methylocaldum szegediense; Candidatus Magnetobacterium casensis*
BGW-G10	Caudovirales	CRISPR	2	2	MG	*Burkholderiales bacterium* GJ-E10*; Comamonadaceae bacterium* H1
BGW-G11	Caudovirales	CRISPR	72	3	MG	*Burkholderia* sp. MR1*; Caballeronia concitans; Pseudomonas aeruginosa*
BGW-G12	*Siphoviridae*	CRISPR	3	3	MS	*Pseudomonas* sp. AAC, ADP, EGD-AKN-5
BGW-G13	Caudovirales	CRISPR	2	2	MG	*Delftia acidovorans; Eikenella* sp. NML130454
BGW-G14	*Myoviridae*	BG	5	4	MG	*Alicyclobacillus macrosporangiidus; Alicyclobacillus shizuokensis; Kyrpidia* sp. EA-1*; Kyrpidia tusciae*
BGW-G15	Viruses	BG	3	3	MG	*Microvirga guangxiensis; Microvirga lotononidis; Caulobacter* sp. K31
BGW-G16	*Siphoviridae*	BG	4	4	MS	*Mycobacterium novocastrense; Mycobacterium rhodesiae; Mycobacterium tusciae; Mycobacterium sphagni*
BGW-G17	Caudovirales	BG	5	4	MS	*Pseudomonas* sp. MT-1, 10B238*; Pseudomonas balearica; Pseudomonas stutzeri; Pseudomonas sagittaria*
BGW-G18	*Siphoviridae*	CRISPR	3	3	MS	*Acinetobacter* sp. 869535, ANC 3862, CIP 102159
BGW-G19	*Siphoviridae*	BG	31	4	MS	*Pseudomonas knackmussii; Pseudomonas aeruginosa; Pseudomonas* sp. HMSC063H08, CCA 1
BGW-G20	Viruses	CRISPR	3	3	MS	*Thermoanaerobacter* sp. X514, X513, X561
BGW-G21	Viruses	CRISPR	2	2	MG	*Proteiniphilum saccharofermentans; Dysgonamonadaceae bacterium*
BGW-G22	*Siphoviridae*	BG	4	3	MS	*Pseudomonas balearica; Pseudomonas stutzeri; Pseudomonas* sp. 10B238
BGW-G23	Caudovirales	BG	27	20	MS	*Acinetobacter lwoffii; Acinetobacter johnsonii; Acinetobacter towneri; Acinetobacter celticus; Acinetobacter gerneri; Acinetobacter indicus; Acinetobacter baumannii; Acinetobacter schindleri; Acinetobacter* sp. ANC 5324, CIP 101934, NIPH 889, NCu2D-2, AR2-3, 51m, HA, WCHA45, ANC 5044, MDS7A, ANC4218, Ver3
BGW-G24	Caudovirales	BG	3	3	MG	*Simplicispira psychrophila; Acidovorax sp. GW101-3H11, KKS102*
BGW-G25	*Siphoviridae*	CRISPR	2	2	MG	*Proteiniphilum saccharofermentans; Dysgonamonadaceae bacterium*
BGW-G26	Viruses	CRISPR	2	2	MS	*Pseudomonas stutzeri; Pseudomonas balearica*
BGW-G27	Caudovirales	BG	2	2	MG	*Thermotalea metallivorans; Sporomusa silvacetica*
BGW-G28	Viruses	BG	2	2	MG	*Riemerella columbina; Salinivirga cyanobacteriivorans*
BGW-G29	Viruses	CRISPR	2	2	MS	*Porphyromonadaceae bacterium KH3R1, NLAE-zl-C104*
BGW-G30	Podoviridae	BG	5	5	MS	*Pseudomonas stutzeri; Pseudomonas balearica; Pseudomonas sagittaria; Pseudomonas* sp. MT-1, 10B238
BGW-G31	Viruses	BG	4	4	MS	*Pseudomonas stutzeri; Pseudomonas balearica; Pseudomonas* sp. MT-1, 10B238
BGW-G32	Caudovirales	CRISPR / BG	9 / 23	22	MS	*Acinetobacter parvus; Acinetobacter haemolyticus; Acinetobacter junii; Acinetobacter lwoffii; Acinetobacter baumannii; Acinetobacter indicus; Acinetobacter towneri; Acinetobacter schindleri; Acinetobacter* sp. CIP 102529, CIP 102143, CIP 102082, WCHA45, ANC5324, AR2-3, 51m, ANC 4218, ANC 5044, HA, NCu2D-2, MDS7A, Ver3, YT-02
BGW-G33	Viruses	BG	4	4	MS	*Pseudomonas stutzeri; Pseudomonas balearica; Pseudomonas* sp. MT-1, 10B238
BGW-G34	*Siphoviridae*	CRISPR	3	3	MG	*Mizugakiibacter sediminis; Hydrocarboniphaga daqingensis; Luteimonas huabeiensis*
BGW-G35	Caudovirales	CRISPR	3	3	MG	*Xanthomonas campestris; Chitiniphilus shinanonensis; Burkholderia plantarii*
BGW-G36	Viruses	CRISPR	5	4	MG	*Pseudomonas aeruginosa; Burkholderia gladioli; Halotalea alkalilenta; Halomonas* sp. 141

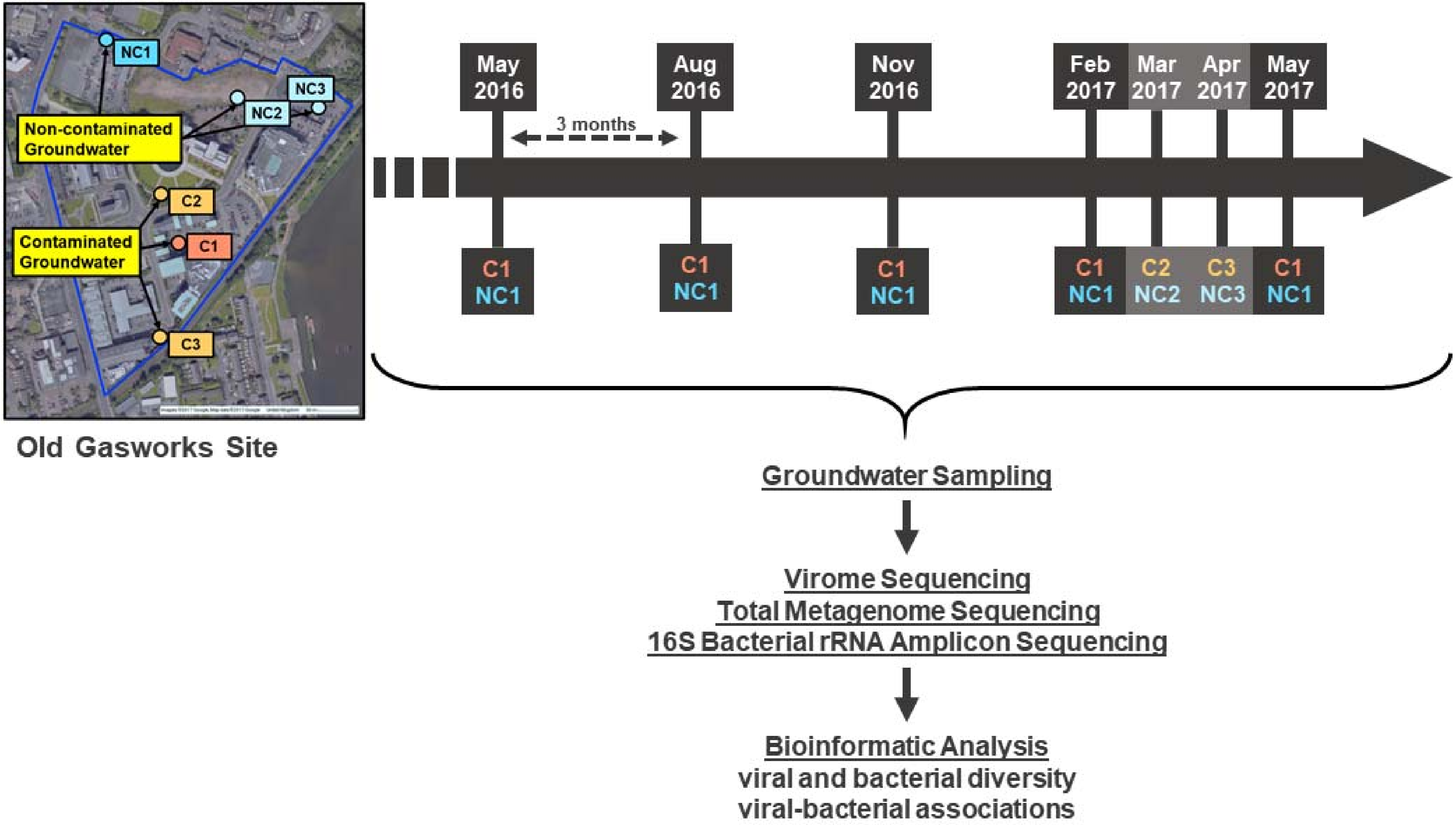

Figure 1. Sampling site and study design. Chronological sampling was done every three months for the period of one year for two sampling stations. Additional sampling was performed at other stations across the site for spatial analysis of microbial community diversity.

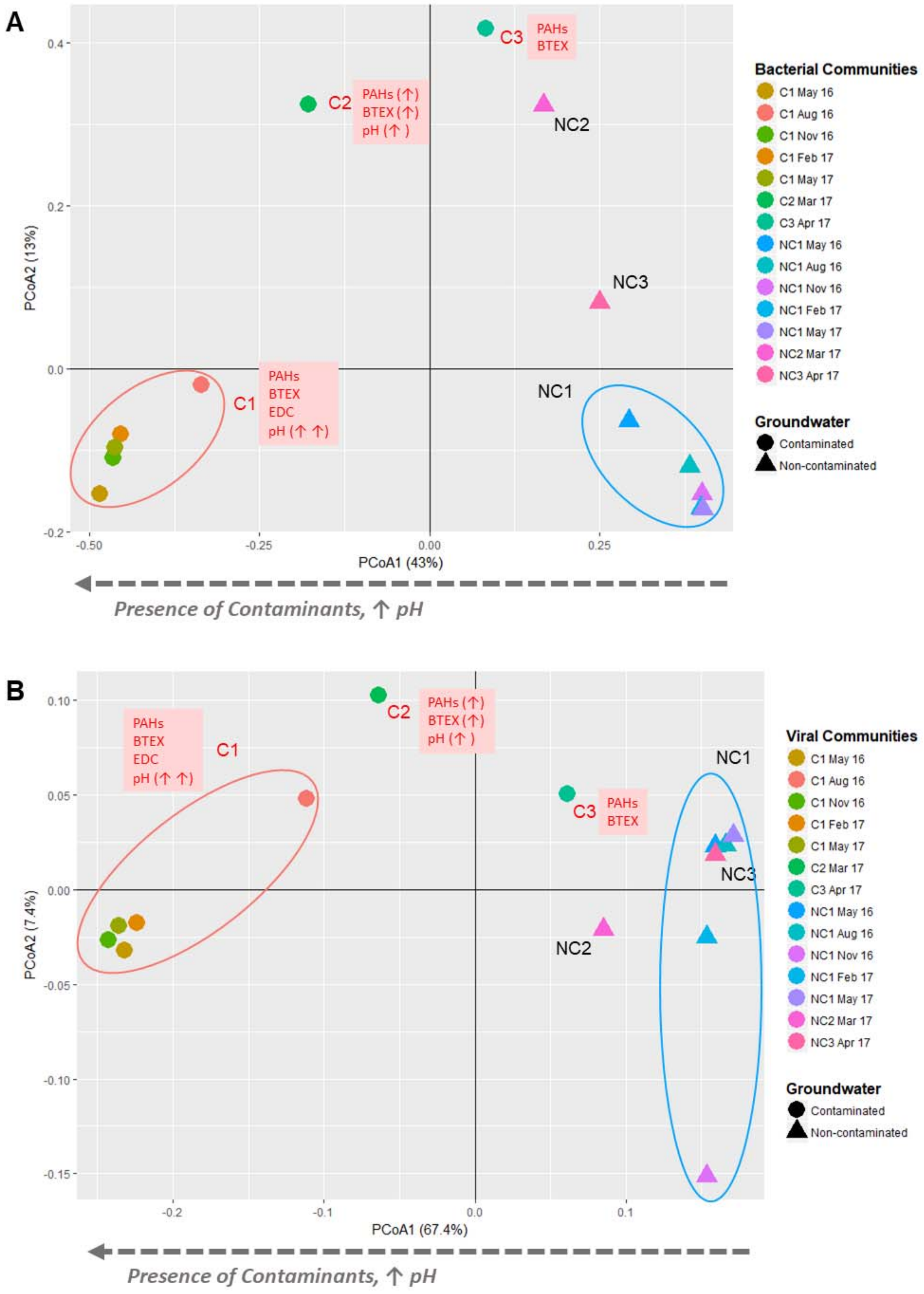

Figure 2. Bacterial (A) and viral (B) cluster analysis of sampled groundwater community diversities. OTUs and virotypes were used to construct PCoA plots based on Bray-Curtis sample dissimilarities.

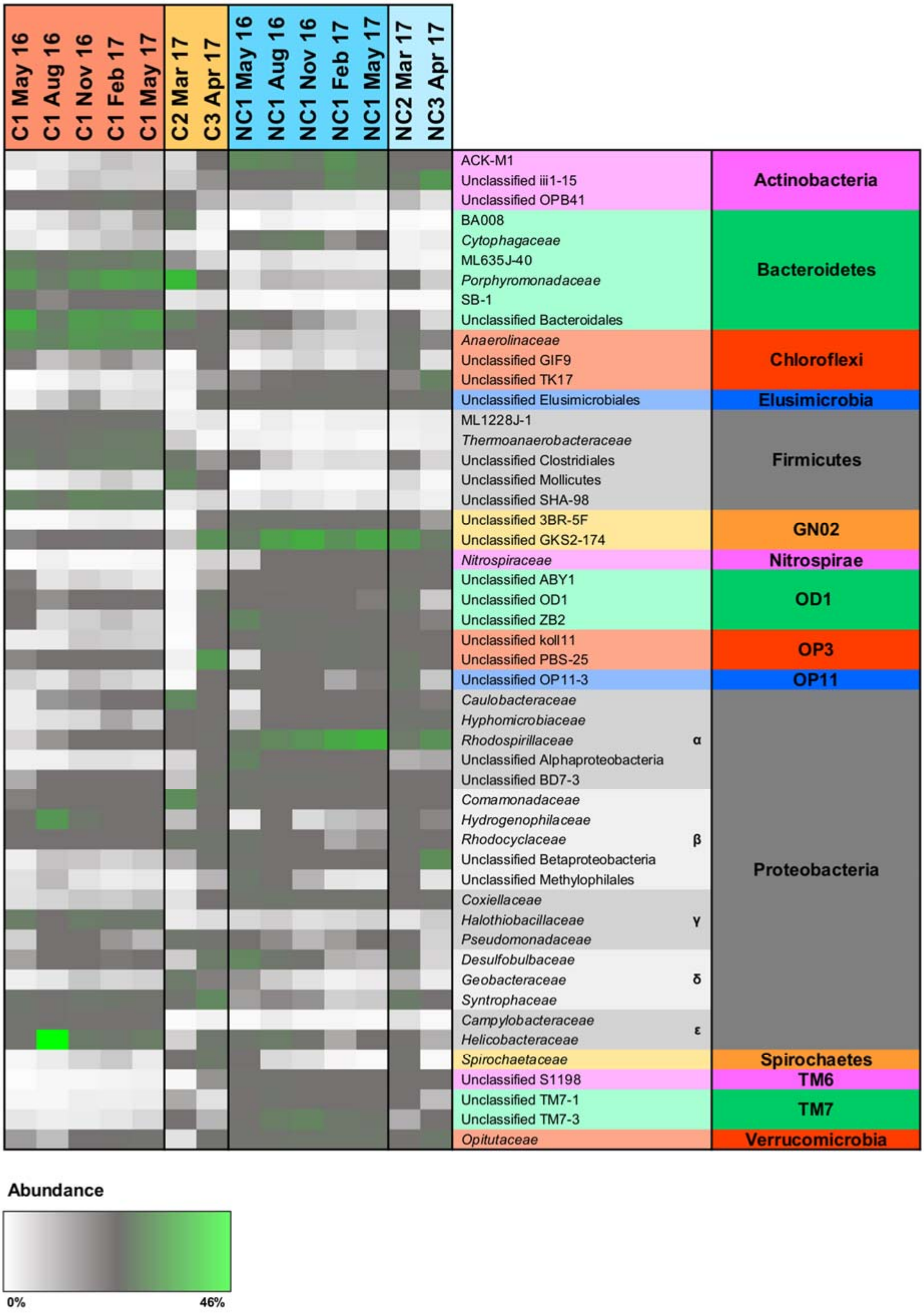

Figure 3. Most abundant bacterial families found in sampled groundwater communities.

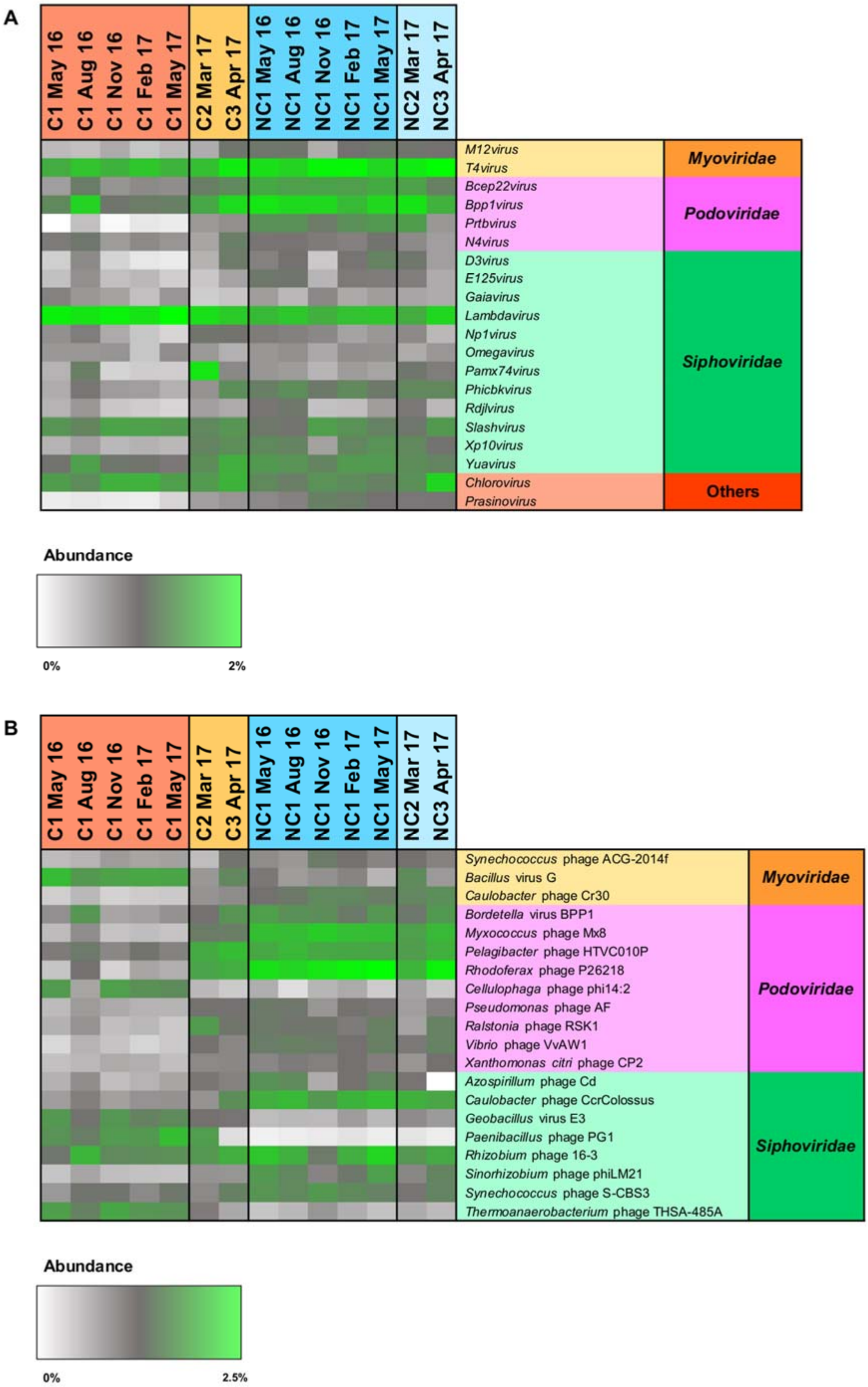

Figure 4. Most abundant viral genera (A) and virotypes (B) found in sampled groundwater communities.

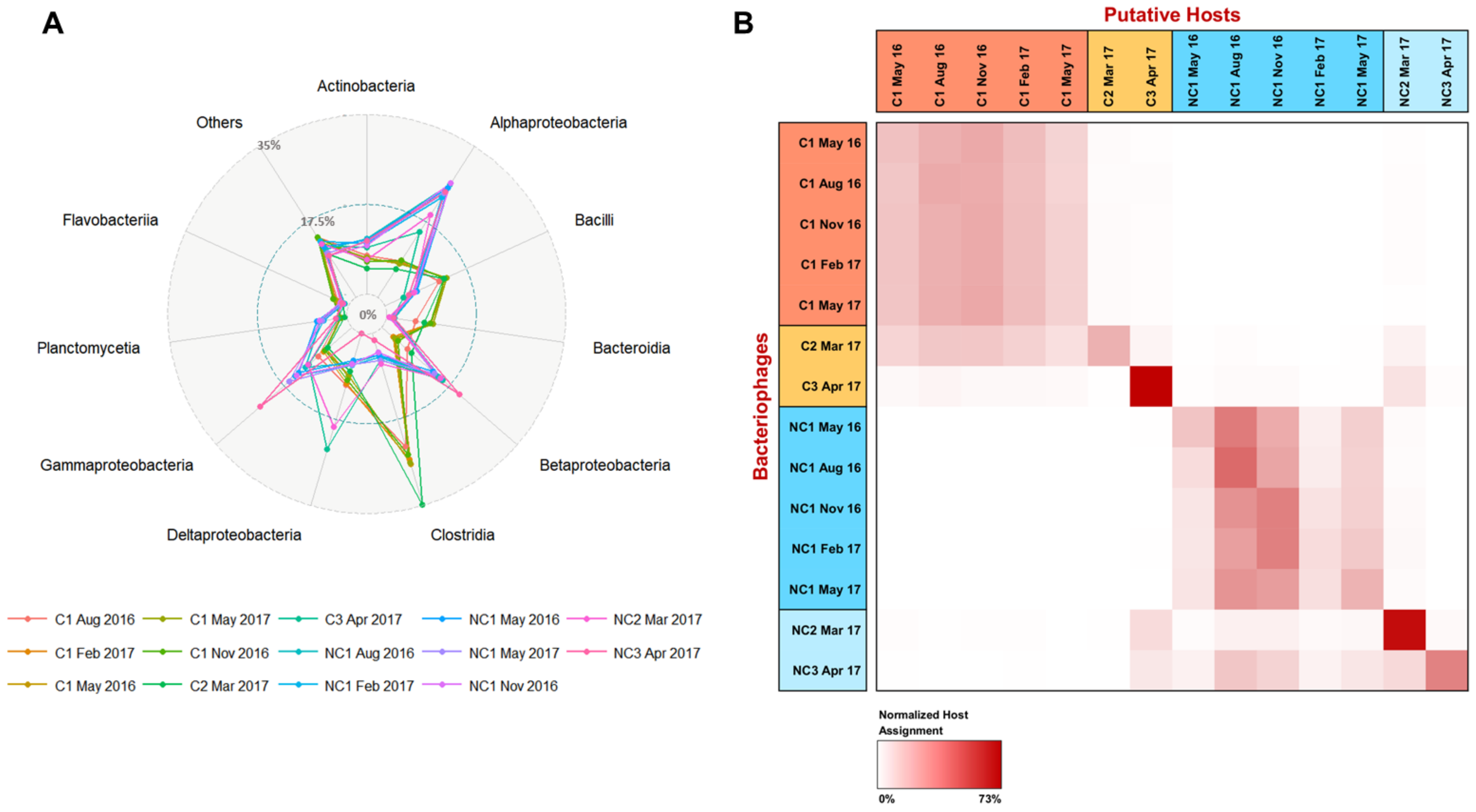

Figure 5. Relative taxonomic abundance of bacteriophage hosts in sampled groundwater communities according to BC homology (A) and relative abundance of putative hosts in sequenced metagenomes (B).

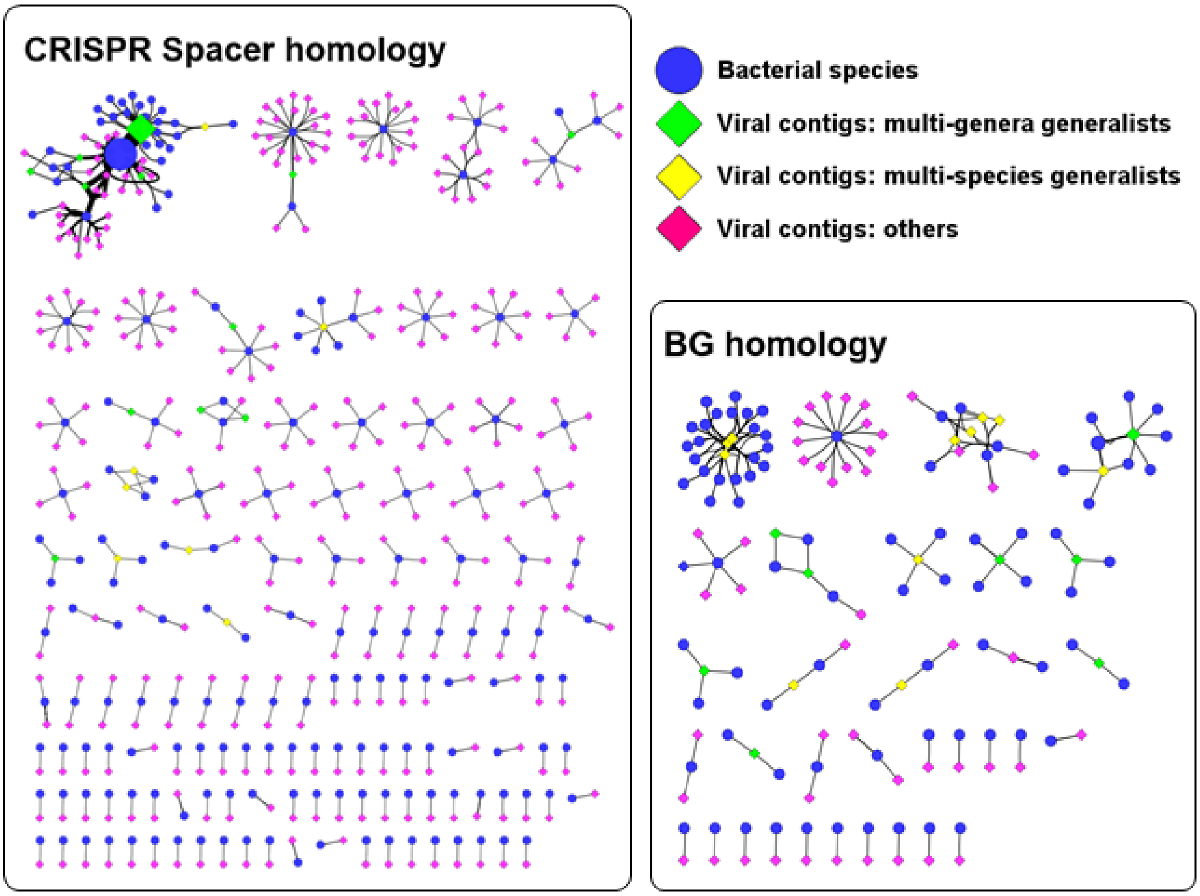
CRISPR Spacer homology
Bacterial species
Viral contigs: multi-genera generalists
Viral contigs: multi-species generalists
Viral contigs: others
BG homology

Figure 6. Viral-host interaction networks based on CRISPR Spacer and BG homology. Viral generalists were classified as multi-species and multi-genera generalists (and above). Size of nodes and edges are proportional to the number of interactions between VCs and bacterial taxa identified.