PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/169717

Please be advised that this information was generated on 2017-08-12 and may be subject to change.
Insights of Phage-Host Interaction in Hypersaline Ecosystem through Metagenomics Analyses

Amir Mohaghegh Motlagh 1, Ananda S. Bhattacharjee 1, Felipe H. Coutinho 2, 3, Bas E. Dutilh 2, 3, 4, Sherwood R. Casjens 5 and Ramesh K. Goel 1*

1 Department of Civil and Environmental Engineering, University of Utah, Salt Lake, UT, USA, 2 Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 3 Radboud Institute for Molecular Life Sciences, Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, Netherlands, 4 Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, Netherlands, 5 Department of Pathology, University of Utah, Salt Lake, UT, USA

Bacteriophages, as the most abundant biological entities on Earth, place significant predation pressure on their hosts. This pressure plays a critical role in the evolution, diversity, and abundance of bacteria. In addition, phages modulate the genetic diversity of prokaryotic communities through the transfer of auxiliary metabolic genes. Various studies have been conducted in diverse ecosystems to understand phage-host interactions and their effects on prokaryote metabolism and community composition. However, hypersaline environments remain among the least studied ecosystems and the interaction between the phages and prokaryotes in these habitats is poorly understood. This study begins to fill this knowledge gap by analyzing bacteriophage-host interactions in the Great Salt Lake, the largest prehistoric hypersaline lake in the Western Hemisphere. Our metagenomics analyses allowed us to comprehensively identify the bacterial and phage communities with Proteobacteria, Firmicutes, and Bacteroidetes as the most dominant bacterial species and Siphoviridae, Myoviridae, and Podoviridae as the most dominant viral families found in the metagenomic sequences. We also characterized interactions between the phage and prokaryotic communities of Great Salt Lake and determined how these interactions possibly influence the community diversity, structure, and biogeochemical cycles. In addition, presence of prophages and their interaction with the prokaryotic host was studied and showed the possibility of prophage induction and subsequent infection of prokaryotic community present in the Great Salt Lake environment under different environmental stress factors. We found that carbon cycle was the most susceptible nutrient cycling pathways to prophage induction in the presence of environmental stresses. This study gives an enhanced snapshot of phage and prokaryote abundance and diversity as well as their interactions in a hypersaline complex ecosystem, which can pave the way for further research studies.

Keywords: bacteriophage, phageome, viromics, phage-host interaction, prophage, phage-host network
INTRODUCTION

Microbial communities and particularly prokaryotes provide diverse ecosystem functions in aquatic habitats. These functions include the cycling of carbon (Jasser et al., 2009; Li et al., 2010), sulfur (Sorokin et al., 2012), nutrients (Kouki et al., 2009; Wang et al., 2013; Motlagh and Goel, 2014), mercury transformation (Chavan et al., 2007; Mitchell et al., 2009; Ramasamy et al., 2012), sulfate reduction (Fortin et al., 2000; Bahr et al., 2005; Faulwetter et al., 2013), and many other biological processes. Therefore, the study of aquatic microbial communities can provide important information about nutrient removal processes, ecosystem functioning and drivers of biodiversity (Gough and Stahl, 2010; Kim et al., 2011).

Bacteriophages outnumber prokaryotic cells in many ecosystems, thus exerting significant predation pressure on their hosts (Williamson et al., 2008). This pressure plays a critical role in the evolution, diversity, and abundance of prokaryotes (Stern and Sorek, 2011). Bacteriophages infect and lyse up to 40% of the prokaryotic population in marine sediments on a daily basis (Middleboe, 2008), resulting in decay of the cell mass and affecting the carbon pool in sediments, which causes more nutrients to be released into the water column. In addition, bacteriophages can influence the genetic diversity of prokaryotic communities in many different ways. Phages selectively kill their hosts, often in a “kill-the-winner” dynamic (Thingstad et al., 2008) in which the most abundant members of a microbial community are the most targeted by phage infection, consequently having their genes temporarily depleted from the genetic pool of a given habitat (Motlagh et al., 2016).

Phages also contribute to the genetic diversification of prokaryotic communities through transduction, a form of phage-mediated horizontal gene transfer (HGT). Through HGT organisms can acquire exogenous DNA from closely or distantly related lineages, and recognition of this process has caused a shift from a diverging “tree of life” view to a “web of life” view of evolution (Olendzenski and Gogarten, 2009; Puigbó et al., 2010). Between 1.6 and 32.6% of genes in prokaryote genomes are estimated to have been acquired through horizontal gene transfer (Koonin et al., 2001), while up to 81% of the genes are estimated to be affected in prokaryotic genomes if the cumulative impact of horizontal gene transfer toward a lineage is also considered (Dagan et al., 2008). Transduction has been studied in various natural environments including freshwater (Kenzaka et al., 2010), marine (Jiang and Paul, 1998), plant-associated (Kidambi et al., 1994), and wastewater systems (Del Casale et al., 2011). However, most previous work has been focused on understanding phage and prokaryote genomes in the transduction process, while the role of transduction in the diversification of prokaryotic genomes and how it affects ecological processes driven by prokaryotes remains poorly understood. This is of special relevance considering the previous research studying the susceptibility of marine microbes to infection by phages isolated from soil, marine sediments, and fresh water and demonstrating that phages move and propagate between major biomes, mediating the transfer of DNA between microbes from very different ecosystems (Sano et al., 2004).

An important component in the study of phage-host interactions is the symbiotic state known as lysogeny, i.e., the stable and non-lytic co-existence of a whole viral genome inside the prokaryotic host genome. Such quiescent phage genomes are called prophages, and they can be integrated into the chromosome of the host bacterium or exist as plasmids. Prophages often account for most of the difference between strains of the same microbial species (Casjens, 2003; Dutilh et al., 2014), and in a “piggyback-the-winner” strategy prophages may become especially important in high microbial density situations (Knowles et al., 2016). Although the integration of temperate phages into the host genome can be beneficial to host and phage (Yosef et al., 2015), these prophages may be induced by a wide range of environmental stress factors such as nutrients (McDaniel and Paul, 2005) or heavy metals (Motlagh et al., 2015) which leads to lysis of the bacterial cell and phage virion release. Moreover, a prophage can dramatically change the phenotype of the host via lysogenic conversion (Paul, 2008). Therefore, studying the presence and role of prophages in natural ecosystems is a critical step toward understanding phage-host interactions.

Studies over the previous decades have greatly increased estimations of the genetic richness of viruses in aquatic environments. Culture-independent techniques such as restriction fragment length polymorphism (RFLP) (Chen et al., 1996) and PCR-based methods such as denaturing-gradient-gel electrophoresis (DGGE) (Short and Suttle, 1999), pulse-field-gel electrophoresis (PFGE) (Bhattacharjee et al., 2015), and hybridization analyses (Richels et al., 1998; Wommack et al., 1999) provided initial snapshots of diversity among prokaryotic communities. However, understanding the taxonomic viral diversity in the environment is challenging since, unlike bacteria, viruses do not carry universally conserved genetic elements (e.g., ribosomal RNA genes of bacteria) that could be used as taxonomic markers to identify all viruses (Rohwer and Edwards, 2002). Metagenomics sequencing approaches have enabled exploration of genetic diversity within environmental samples, thus evading limitations associated with conventional culture-dependent microbiology methods. In addition, by targeting total nucleic acids, shotgun metagenomics allows functional and taxonomic characterization of samples without the need for prior knowledge of the microbial types present in the studied environment.

Previous studies have demonstrated the effect of phages and their infection processes on the sediment microbiome in different environments such as wetland (Jackson and Jackson, 2008), marine (Rohwer and Thurber, 2009), and freshwater (Short and Suttle, 2005). However, hypersaline environments are among the least studied ecosystems, and so there is a lack of knowledge about the interaction between prokaryotic communities and bacteriophages in such habitats. The Great Salt Lake (GSL) is the largest prehistoric hypersaline lake in the Western Hemisphere, having a unique geology that gives rise to special ecologic and economic domains of relevance in its various uses as both a migratory bird habitat and a source of brine shrimp, trace elements, and other minerals. In the present study, we employed metagenomics analysis to determine the diversity of phages and prophages in the GSL, as well as to...
better understand the effects of bacteriophages in defining the diversity and interactions among prokaryotic communities and their biogeochemical cycles.

**MATERIALS AND METHODS**

**Site Description and Sediment Sampling**

The Great Salt Lake (GSL) is a terminal lake that represents a complex hypersaline ecosystem with a salinity gradient between 6 and 28% that is contaminated with, for example, mercury, selenium, and nutrients (Williams, 2002). The lake is fed by the Weber, Bear, and Jordan rivers. These rivers carry more than 1.1 million tons of salts annually into the lake (Rafferty, 2011), which transformed GSL into one of the most saline bodies of water in the world. The total dissolved mineral accumulation in the lake basin is estimated at 5 billion tons, consisting of mainly sodium, and chloride ions, though sulfate, magnesium, and potassium are also abundant (Stephens, 1998). The sediment sample was collected 6 miles west of Antelope Island (40°53′51.3″ N, 112°21′00″ W) in the South Arm of the Great Salt Lake (Figure S1). Physicochemical data in the water column were obtained using a sounder equipped with pH, specific conductivity, water temperature, depth, and dissolved oxygen sensors. The sampling was carried out in June 2014 with the collaboration of the United States Geological Survey (USGS). Sediment sample was collected at a depth of 27 ft from the deep brine layer at the surface of the sediment with a stainless steel box corer (Wildco, FL). The sediment sample was immediately stored on ice and shipped to the laboratory for analysis.

**Characteristics of the Environmental Sample**

For nutrient determination, sediment sample was centrifuged at 2,000 × g for 10 min to extract interstitial and pore water. The pore water was filtered with 0.45 µm mixed cellulose hydrophilic filter paper (Millipore, MA) and ammonia (NH₃), nitrate (NO₃⁻), nitrite (NO₂⁻), total nitrogen (TN), total phosphorus (TP), and total organic carbon (TOC) were quantified using HACH methods HACH 8155, TNT839, TNT835, TNT826, TNT843, and HACH 10128, respectively, according to the manufacturer's instructions. In order to ensure that the high salinity of the samples did not interfere with the measurement procedures and reading the spectrophotometer, control samples with known concentration of ammonia, nitrite, nitrate, and phosphorus were also used alongside each sample. Total solids (TS) and volatile solids (VS) were also measured by drying the sediment sample at 103°C for 24 h followed by igniting at 550°C for 2 h according to EPA Method 1684 (USEPA, 2001). In addition, to study the biogeochemistry of the sediment sample, trace metal concentrations (Mn, Fe, Co, Ni, Cu, Zn, Se, Mo, and Pb) were measured by using 15 g of sediment, centrifuging at 5,000 × g for 10 min and filtering the supernatant with 0.22 µm hydrophilic filter and then analyzed using inductively coupled plasma mass spectrometry (ICP-MS, Agilent 7500ce, CA).

**Microbial Nucleic Acid Extraction**

Five grams of sediment were re-suspended in sterilized 10 ml of PBS buffer and vortexed for 5 min to homogenize the sediment. The homogenized sample was then centrifuged for 5 min at 2,000 × g and 250 µl of clear supernatant devoid of any cells was discarded. The pellet was re-suspended in 1 ml of PBS buffer and 500 µl of sample was used for DNA extraction with different methods including phenol-chloroform DNA extraction (Köchl et al., 2005), cetyl trimethyl ammonium bromide (CTAB) extraction (Zhou et al., 1996), PowerSoil® DNA isolation kit (MoBio laboratories, CA), and PowerMax™ Soil DNA isolation kit (MoBio laboratories, CA) following the manufacturers' procedures. In order to minimize the extraction biases, the extracted DNA was pooled and the quantity and quality were verified on a NanoDrop ND 2,000 spectrophotometer (Thermo Scientific, USA) at 260 and 280 nm. The DNA purity and quantity were also verified on 1.2% agarose gel prior to high-throughput sequencing.

**Bacteriophage Isolation and Nucleic Acid Extraction**

Free phages were extracted by re-suspending 250 g of sediment in 3× volume of filter sterilized 1% (w/v) potassium citrate buffer (10 g/l potassium citrate, 1.44 g/l of disodium phosphate, 0.24 g/l of monopotassium phosphate, pH 7). The sediment with potassium citrate buffer was placed on a shaker on ice overnight to suspend free phages in solution. Thereafter, the mixture was centrifuged at 8,000 rpm on an Avanti J-E centrifuge (Beckman Coulter, CA) for 30 min to pellet the bacterial debris followed by centrifugation of the supernatant at 9,000 rpm for 12 h to concentrate the phages. Following overnight centrifugation, the pellet was re-suspended in SMG buffer (5.8 g/l sodium chloride, 2 g/l magnesium sulfate, 5 ml/l of 5% (w/v) gelatin, 50 ml/l of 1M Tris–Cl, pH 7.5) and filtered with 0.22 µm pore size filter paper (Millipore Co., MA) to remove any residual bacterial cells and sediment debris.

The phage particles were purified using cesium chloride (CsCl) gradients at 1.35–1.6 g/ml density by isopycnic centrifugation at 35,000 rpm for 3 h (Angly et al., 2006). This process was carried out twice to ensure the removal of any residual bacterial cell debris and guarantee the purity of viral particles. Before DNA extraction from the virions, the sample was subjected to DNase treatment with RNase-free DNase I (Thermo Scientific, CA) at 37°C for 30 min. This step digested any free DNAs in the sample. Subsequent phage DNA extraction was performed with a phage DNA isolation kit (Norgen Biotek Corp., Canada). In addition, to confirm the absence of microbial DNA contamination in the phage DNA extracts, PCR amplification of the hypervariable V4–V9 region was tested using 515F/942R universal 16S rRNA gene primer set (Diemer and Stedman, 2012). Aliquots of the amplification products were electrophoresed in a 1.2% agarose gel stained with ethidium bromide (10 µg/ml) and visualized under UV illumination.
DNA Library Preparation for Sequencing

Amplification methods such as random amplified shotgun library (RASL), linker-amplified shotgun library (LASL), and multiple displacement amplification (MDA) are usually implemented to increase the nucleic acid yield and obtain a sufficient quantity of DNA for sequencing. However, these methods can result in quantitative biases such as selective amplification of single-stranded DNA (ssDNA) viruses' genomes (Kim et al., 2008) and production of artifacts such as chimeras (Lasken and Stockwell, 2007). Therefore, we refrained from using any amplification process in order to minimize biases for the phage metagenomics and analysis of the phageome data.

Library construction was performed using the Epicentre EpiGnome Methyl-Seq Kit as described below. Briefly, genomic DNA (~50 ng) was heat denatured and hybridized with oligonucleotides consisting of random hexamers linked to Illumina P5 adapter sequences. Strand replication was accomplished using EpiGnome polymerase. Double-stranded DNA was heat denatured to enable ligation of the EpiGnome Terminal Tagging Oligo which adds Illumina P7 adapter sequence to the 3' end of the replicated strand. Adapter-ligated DNA molecules were enriched by 10 cycles of PCR and the amplified library was subsequently purified using Agencourt AMPure XP beads (Beckman Coulter Genomics, CA). The concentration of the library was measured using the Qubit dsDNA HS Assay (Invitrogen, CA) and an aliquot of the library was resolved on an Agilent 2,200 Tape Station using a D1000 assay to define the size distribution of the sequencing library. Libraries were adjusted to a concentration of approximately 10 nM and quantitative PCR was performed using the Kapa Library Quant Kit (Kapa Biosystems, MA) to calculate the molarity of adapter-ligated DNA molecules. The concentration was further adjusted following qPCR to prepare the library for Illumina sequence analysis and the samples were sequenced on an Illumina MiSeq Benchtop DNA sequencer (Illumina, CA) with 300-cycles paired-end at HCI Core Facility of University of Utah.

Metagenomic Analysis

Paired-end raw reads were interleaved, quality filtered and trimmed using CLC Genomics Workbench v.7.0.4 (CLC Bio, Denmark) with a threshold of 100 bp as the maximum length of read and a Phred score of 28. The trimmed reads were de novo assembled using CLC Genomics Workbench v.7.0.4 with the following criteria: word size of 20 bp, automatic bubble size of 50 bp, and minimum contig length of 500 bp. Identification of open reading frames (ORFs) and gene prediction were performed using MetaGeneMark v.2.8 (Zhu et al., 2010) followed by gene annotation using the RPSBLAST program (Altschul et al., 1990) on the clusters of orthologous group (COG) prokaryotic protein database (Tatusov et al., 2000). Statistical over-representation of annotated COG genes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was determined by pairwise comparisons of each metagenomic sample using Fisher's exact test, with confidence intervals at 99% significance. Besides COG analysis, the predicted genes were categorized into SEED Subsystems to provide consistent and accurate genome annotations (Overbeek et al., 2014) through MG-RAST (metagenomics rapid annotation using subsystem technology) pipeline to identify all of the encoded proteins and their functions in the host metabolism. The enzymes were compared against SEED Subsystems using a maximum e-value of 1E-5, a minimum identity of 60%, and a minimum alignment length of 100 measured in amino acids for protein databases. The minimum alignment length of 100 amino acids (equal to 300 bps) was chosen as the criterion to specify the possible functional genes (not only partial genes) that are transferred from prokaryotes to the phages.

In order to understand the phage-host interactions, the GC contents of phage and prokaryote contigs were calculated and compared with contig percentage shown in a histogram. The occurrence of each single tetranucleotide in the phage and prokaryote contigs was calculated using JSpecies v.1.2.1 (Richter and Rossello-Mora, 2009). In addition, the calculated GC content of phage and prokaryote contigs were graphed along with the scaffold length and contig coverage using the ggplot package in R v.3.0.1. The sequences are deposited in MG-RAST repository with accession number mgm4582011.3 and mgm4582012.3 for bacteria and phage sequences, respectively.

Analysis of Prokaryotic Metagenome

The taxonomic composition of the prokaryotic metagenome was determined by comparison of the assembled contigs against the NCBI non-redundant (nr) nucleotide database using tBLASTx v. 2.2.28 program (Altschul et al., 1990) with 1E-5 e-value cut-off. Relative abundance of taxa was assessed on the basis of average counts of mapped reads on the BLAST annotated contigs. Therefore, for the reads that were mapped to contigs, every blast hit was multiplied by the average depth of the reads on each contig. Eventually, significant hits to GenBank entries were recorded in a BLAST output file and imported on MEGAN v.5 (Huson et al., 2007) for interpretation with the lowest common ancestor (LCA) method followed by visualization in iTOL online tool (Letunic and Bork, 2007). For the MEGAN analysis, default parameters for each metagenome were selected as follows: minimum support of 5, minimum score of 50, top percent of 10, win-score of 0, and minimum complexity of 0.44.

The affinities of the sequences for known metabolic functions were also annotated using BLASTx with cut-off e-value of 1E-5 against SEED subsystems v.2.0 (Overbeek et al., 2005) using MG-RAST server v.3.3 (Meyer et al., 2008) and KEGG metabolic pathways (Kanehisa et al., 2008). In addition, to understand the role of prokaryotic communities in nutrient cycles, the identified prokaryotes were compared manually with a curated database on functional genes (FunGene) involved in various biogeochemical cycles (Fish et al., 2013).

Analysis of Phage Metagenome

In addition to initial viral particle purification step prior to sequencing and DNase treatment of the purified phage prior to DNA extraction, in order to exclude any bacterial contamination from the phage contigs, genes encoding the 5S, 16S, and 23S rRNAs (from prokaryotic genomes) were identified in phage...
contigs using hmm_rRNA on WebMGA server based on an HMM search using HMMER v.3 (Finn et al., 2011).

Since analysis of the phage metagenome was also performed on contigs instead of reads, the relative abundance of contigs was corrected by the average depth of the reads on each contig as mentioned earlier. For phage taxonomic analysis, the filtered phage contigs were compared against NCBI RefSeq viral genome database by using tBLASTx v. 2.2.28 program with e-value of 1E-5 cut-off. The generated BLAST output file was imported on MEGAN v.5 for interpretation by lowest common ancestor (LCA) method using the default parameters as mentioned above (results not shown). Instead, the abundance of phages from the GSL was explored with respect to metagenomes of marine, freshwater, and lake sediment samples. Raw reads of aquatic viromes available on Metavir (Roux et al., 2011) were mapped to assembled phage contigs with length longer than 20 kbp as putative complete phages using Bowtie v.2 (Langmead and Salzberg, 2012). Contig abundances were obtained by counting the number of reads mapped to each contig and correcting by contig length. The obtained abundance matrix with identity percentages was normalized and plotted as a heatmap in which samples and contigs were clustered according to their Euclidian distance using R v.3.0.1.

Prophage Identification and Analysis
To achieve a more detailed understanding of the phage-host interactions, mobile genetic elements including prophages, plasmids, and transposons were investigated in the prokaryote contigs. Following gene prediction in the prokaryote contigs, integrated prophages were detected using Prophinder (Lima-Mendez et al., 2008) to identify mobile genetic elements. Simultaneously, in order to relate the prophages to their host, prokaryote contigs were compared against ACLAME (A CLAssification of Mobile genetic Elements) database (Leplae et al., 2004) using a BLASTn search with e-value 1E-5 cut-off, followed by extraction of the prophage hit regions (excluding viruses and plasmids) from the BLAST output results. These prophage regions were compared with prophages found using Prophinder and the shared prophages with their corresponding prokaryotic hosts were extracted to generate the taxonomic cladogram using MEGAN with default parameters mentioned earlier. In addition, nutrient cycling of the prokaryotic hosts was manually annotated based on the KEGG database and used to generate the prophage-host interaction network using iTOL online tool.

Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)
Clustered regularly interspaced short palindromic repeat (CRISPR) is an anti-viral mechanism in archaea and bacteria wherein genomic sequences from the predatory viruses are integrated in the host genome providing immunity to these viruses (Horvath and Barrangou, 2010). Therefore, CRISPR spacer sequences can provide a direct link between viruses and their prokaryotic hosts (Kunin et al., 2008; Heidelberg et al., 2009; Anderson et al., 2011) and CRISPR spacers may be viewed as a database of fragments derived from phage and plasmid genomes. In addition, phage genome can acquire some prokaryotic genes during an infection event of the host and therefore, homology of phage and prokaryotic genes can indicate phage-host associations. In our study, the assembled bacteria contigs were compared against phage contigs as the database by using tBLASTx v. 2.2.28 setting a 1E-5 e-value cut-off to find their homology. Afterward, all CRISPR arrays in the prokaryote contigs with homology to phage contigs were identified using CRISPR Recognition Tool (CRT) v.1.1 (Bland et al., 2007). Putative protospacer targets were also identified using CRISPRTarget (Biswas et al., 2013), following a BLASTn search of the spacer input against ACLAME, GenBank Environmental, GenBank Phage, and RefSeq plasmid, RefSeq viral databases with the default settings.

RESULTS
Characteristics of the Environmental Sample
Following sample collection from 27-ft deep brine layer, the environmental quality parameters, nutrients, and metals of the sediment sample were measured as summarized in Table 1. Measurement of nutrient and metal concentrations in the sediment can be an indicator of microbial activities involved in different nutrient cycles. Salinity was measured as the total salt concentration, comprised mostly of Na⁺ and Cl⁻ ions. The measured salinity of the GSL was 15% (150 g/l), which is approximately 5 times the average salinity of the ocean and 300 times the average salinity of fresh water (Kerr et al., 2003). In addition, significantly low dissolved oxygen levels in the deep brine layer and high concentration of total phosphorus (TP) and total nitrogen (TN) make the water of the Great Salt Lake an extreme environment, which selects for a microbiota adapted to high salt concentrations.

Total organic carbon (TOC) concentration was also measured in the sediment sample, and it showed a high carbon amount of 20.8 mg.g⁻¹ (2%) compared to the average TOC of 0.5% in the deep ocean (Seiter et al., 2004). Moreover, a considerable amount of ammonia (NH₃) was measured in the sediment sample compared to the EPA aquatic life water quality criteria for ammonia (Unites States Environmental Protection Agency, Office of Water, 2013), which results in a very low total organic carbon to total nitrogen (C/N ratio) that can mobilize nitrogen in the sediment due to its excess amount.

Prokaryote and Phage Metagenomes Contig Assembly
The high-throughput sequencing generated 15.1 million reads for each of the bacteria and phage DNA samples. Following quality and length control, 1.26 million (8.34%) and 1.27 million (8.39%) reads from the bacteria and phage samples were removed, respectively. The filtered and interleaved paired-end reads were de novo assembled using CLC Genomics Workbench resulting in 40,223 contigs with average length of 1,575 bp, N50 of 1,573 bp, and average coverage of 20× for bacteria reads; and 45,689 contigs with average length of 1,595 bp, N50 value of 1,615 bp,
and average coverage of $13 \times$ for phage reads. Average coverage of the contigs was calculated by multiplying the read quantity by the average read length and dividing by the average contig length. These contigs contained 51.7 and 43.9% of the shotgun metagenomics reads in the bacterial and phage metagenomes, respectively.

Although typical prokaryotic genes (such as 16S ribosomal RNA) were never detected in the phage genomic DNA prior to sequencing, to ensure the absence of any cellular DNA in the phage sequences, genes encoding the 5S, 16S, and 23S rRNAs were identified in phage contigs. A total of 0.1% of the phage sequences, genes encoding the 5S, 16S, and 23S rRNAs respectively.

### Microbial Community Structure

**Prokaryotic Community Structure and Diversity**

Microbial communities and specifically prokaryotes in natural ecosystems are involved in nutrient mobilization and regeneration, primary production, and energy fluxes. In addition, the prokaryotic populations in the sediment play key roles in biogeochemical cycles and their infection by phages can affect various nutrient cycles in which these organisms are involved. Investigating microbial diversity is therefore essential to understand the ecology and microbial interactions of Great Salt Lake.

For a better understanding of prokaryote diversity, following tBLASTx analysis against NCBI non-redundant (nr) nucleotide database with 1E-5 e-value cut-off, 72.3% of the contigs matched the bacterial domain, while archaea (1%), eukaryotes (6.4%), viruses (13.8%), and unassigned with no annotation (6.4%) comprised the rest of the contigs. The prokaryotic taxonomic rRNA analysis revealed that Proteobacteria (46.2%), Firmicutes (20.3%), Bacteroidetes (9%), Actinobacteria (5.3%), Chloroflexi (5.2%), Cyanobacteria (2.7%), and Planctomycetes (1.7%) were the most dominant bacterial phyla present in the sediment sample. Prokaryote taxonomic diversity and population found in the prokaryote contigs are presented in Figure S2, Tables S1, S4.

**Bacteriophage Community Population and Diversity**

Since unassembled reads in viral metagenomes are often characterized by a majority of unknown sequences that have no homologs in the database (Mokili et al., 2012), assembled reads present in phage contigs were used for taxonomic characterization to facilitate their annotation. Classification of viral metagenome contigs using lowest common ancestor (LCA)
Mohaghegh Motlagh et al. Metagenomics of Phage-Host Interactions in Hypersaline Sediments

**FIGURE 1** | (A) GC content distribution of bacteria and phage contigs. (B) Standardized (z-score) tetramer word frequencies of phage and bacteria contigs with the regression line with shaded confidence region.

**FIGURE 2** | Overlapped bacteria (open circles) and phage (solid circles) showing contigs with similar GC content and coverage suggesting that phages were well-adapted to the codon usage of their prokaryotic hosts. All circles represent contigs, scaled by their length and colored by GC content. The regression line with shaded confidence region is shown and arrows are illustrating some of the overlapped contigs.

Analysis with the RefSeq viral database revealed *Siphoviridae* (32%), *Myoviridae* (24%), and *Podoviridae* (13%) as the most dominant viral families in the sediment sample. In addition to these dominant phage families, other phages such as unclassified dsDNA phages (6%), unclassified archaeal dsDNA virus (6%), and unclassified *Caudovirales* (3%) were among matched hits in the phage contigs. Among *Siphoviridae* viral family, unclassified lambda-like phages was the most dominant phage with 6% among the family and 2% in the entire phage community. In addition, analysis of *Myoviridae* family showed that unclassified T4-like phages and T4-like phages in the subfamily of *Tevenvirinae* was composing 7 and 8% of the *Myoviridae* family, respectively. Furthermore, *Cellulophaga* phage was among the most dominant phages in *Podoviridae* comprising 8% of the family and 1% of total phage community.
Despite the growing knowledge for the ecological role played by bacteriophages and an increase in phage sequence information in public databases, little is known about their dynamics in natural ecosystems. It is therefore imperative to be able to classify these organisms in order to determine their dynamics. As a result, this study developed a higher resolution insight into the viral biogeography by mapping assembled phage contigs with length longer than 20 kbp as putative complete phages from the GSL to phage sequences from freshwater lakes, lake sediments, and marine ecosystems to determine the biogeographical distribution patterns of these phage genome fragments (Figure 3). In the generated heatmap, cell color represents the relative percentage of identity between GSL contigs and other publicly available phage metagenomes. From the high resolution abundance heatmap, it was interesting to observe that these novel unclassified Caudovirales GSL phages tend to be also abundant in marine environments as well despite Great Salt Lake being a hypersaline land-locked lake.

**Predicted Gene Abundance and Analysis**
Gene prediction and ORF identification in phage and bacterial contigs were performed to identify possible events of horizontal
gene transfer between the phages and their prokaryotic hosts. GSL metagenome predicted genes were classified into one of 22 Clusters of Orthologous Groups (COG) functional categories according to best hit classification through RPS-BLAST against the COG reference database. Out of the 122,016 bacterial and 102,594 viral predicted genes, 26,319 and 24,495 could be annotated with e-value cut-off of 1E-3. Figure 4 shows the frequency of phage and bacterial genes in various COG classes. The category J (translation, ribosomal structure, and biogenesis) and category Q (secondary metabolites biosynthesis, transport, and catabolism), were significantly different for phage and bacterial predicted genes. Both categories have down-regulated genes (Chhabra et al., 2006) suggesting that some of the genes in these categories are responsible for the expression inhibition of host genes. In addition, genes classified in category Q have a role in response to environmental changes (Makino et al., 2003), which were significantly less available in the bacterial genes.

**Prophage Identification**

Identification of prophages in the bacterial contigs was applied to determine phage-host interactions in the natural ecosystem. Using Prophinder, a total of 69 prophages (33 intact and 36 questionable prophages) were found in the bacterial contigs. From these 69 total identified prophages, 37 prophages were joined to recognizable host DNA in the bacterial contigs and so were assigned to their prokaryotic hosts. Prokaryotic host DNAs were not found in the bacterial contigs for the other 32 putative prophages, and they were therefore excluded from further analysis. The nutrient cycling processes of the propharyotic hosts were based on the KEGG database and were manually annotated. The taxonomy of prophages in the bacterial contigs was plotted on a circular cladogram relating the prophages and their prokaryotic hosts involved in different nutrient cycles (Figure 5), and it showed that carbon cycle was the most susceptible nutrient cycling pathway to prophage induction in the presence of environmental stresses. The numbers of different nutrient cycling pathways that can be affected by prophage induction are plotted in the Venn diagram (Figure S3) and the prophage-host interaction showing the identified prophage, their prokaryotic hosts are presented in Table S2 in detail.

**Crisprs Identification**

One limitation of metagenomics analyses is that when bacteria and viruses are sequenced together, it may be difficult to
distinguish between the two types of sequences. Conversely, when phages are isolated and sequenced directly, it is challenging to identify the specific hosts from the viral sequence (Edwards et al., 2016). However, such information is critical for the analysis of the relationship between phage and host. In order to approach this problem, similarities between bacterial CRISPR spacers and the infecting mobile elements were analyzed to study these interactions in the natural ecosystem.

Initially, in order to associate the CRISPRs with phage sequences, the bacterial contigs as query were compared with phage contigs as database using BLASTn with an e-value 1E-5 cut-off. A total of 6,025 bacterial contigs with more than 70% identity in sequence with phage contigs were extracted.
Using CRISPR Recognition Tool (CRT), 96 CRISPR arrays were recognized from all bacterial contigs with an average repeat length of 25 bp and spacer length of 40 bp. These identified CRISPR arrays were explored to determine targets of CRISPR RNAs using CRISPRTarget. Interestingly, the identified genome spacers showed similarity with eight of the protospacers on the marine phage metagenome from TARA database (Sunagawa et al., 2015). These were shown in the previous section to have considerable homology with the phage contigs in the GSL sample. This connection indicates that previous infections of GSL prokaryotic hosts with relatives of these phages must have occurred.

DISCUSSION

In this study, we used phage and bacterial metagenome sequencing to better understand the microbial interactions between the phage and their prokaryotic hosts. The nutrient concentrations in the sediment were measured to explore the microbial activities involved in different nutrient cycles. As mentioned earlier, considerable concentration of organic carbon was measured in the sediment sample. Organic carbon deposited in sediment is mainly derived from freshly deposited plant litter and decomposed forms such as humus (Kristensen and Holmer, 2001). In addition, bacterial growth and metabolic activities of methanotrophs, methanogens, and photosynthetic phytoplanktons contribute to the high concentration of TOC in sediment sample through fixation of CO₂. As discussed on prokaryotic community structure, our metagenomic analysis shows the presence of halophilic and halotolerant methaneproducing and methanotroph bacteria in the GSL sediments supporting the presence of such bacterial carbon sources in the sediment. Furthermore, high concentrations of nutrients can increase phytoplankton growth and consequently increase of debris sedimentation. As a consequence of the increased productivity that results in oxygen depletion, sediments in the anoxic environment of the lake will contain a larger amount of organic matter (Lazar et al., 2012). The nitrifiers including Nitrospirae, Nitrosomonas, and Nitrobacter, and denitrifiers including Bacillus, Pseudomonas, Clostridium, Halomonas, and Rhodobacter were among the most abundant phyla found in the prokaryotic contigs, which considerably affect the nitrogen cycle. In addition, ammonia concentrations following the appearance of brine fly larvae and brine shrimp in the GSL in early summer (Wurtsbaugh, 1992) when algal populations decline due to grazing and nitrogenous wastes from the invertebrates increase. Furthermore, ammonium (NH₄⁺) can arise from decomposition of organic matter in the water column or from diffusion of ammonium from the anaerobic sediment layer into the water column (Chowdhury and Bakri, 2006). Furthermore, presence of nitrogen fixing bacterial family such as Beijerinckiacae as well as dissimilatory nitrate reduction to ammonium (DNRA) bacterial family such as Geobacteraceae in the analyzed bacterial metagenome were contributing to increase the level of ammonia in the sediment sample. The levels of trace metals observed at the deep brine layer suggest that the sediment may be a hotspot for nutrient cycles. Many enzymatic pathways of the nitrogen, carbon, and sulfur cycles require cofactors which contain iron or molybdenum. In particular, iron is a necessary requirement and an integral component of many enzymes involved in photosynthesis, electron transport, and nutrient acquisition (Geider and La Roche, 1994). Furthermore, molybdenum is a required component of enzymes involved in nitrogen cycling pathways such as nitrate assimilation and nitrogen fixation (Glass et al., 2015). For instance, nitrogenase, a critical multi-component enzyme in nitrogen fixation typically consists of iron and molybdenum-iron containing subunits (Howard and Rees, 1996). As discussed in the following gene analysis, these critical enzymes involved in various nutrient cycles were found in both prokaryote and phage metagenomes, which suggests a phage-mediated gene transfer process.

More than 450 different genera were classified in the sediment sample, while the most abundant genera, in order of abundance, are Desulfovococcus, Halanaerobium, Desulfovibrio, Desulfatibacillum, and Streptomyces, which belong to taxa involved in various biogeochemical cycles. Desulfovococcus plays an important role in cycles of sulfur compounds in sea water (Das et al., 2006). Halanaerobium is an obligatory anaerobic halophile, and several strains of Halanaerobium such as H. praevales were previously isolated from GSL and showed complex organic matter fermentation and production of intermediary metabolites for other trophic groups such as sulfate-reducing and methanogenic bacteria (Ivanova et al., 2011). Presence of Desulfovibrio and Desulfatibacillum, as sulfur-reducing anaerobic bacterial genera, was consistent with the anoxic condition of the brine layer sediment. Streptomyces is an important organism in carbon recycling that has a crucial role in the environment since it carries out a broad range of metabolic processes such as degradation of insoluble biological material including lignocellulose and chitin (Bentley et al., 2002). Proteobacteria of the genus Rhodobacter are metabolically versatile, capable of aerobic and anaerobic respiration, and anoxygenic photosynthesis when grown anaerobically in the light (Han et al., 2004).

Caudovirales are tailed bacteriophages, which are found to dominate in marine and other aquatic environments (Weinbauer, 2004; Winter et al., 2014). In a recent study, Antunes et al. (2015) studied viral communities in deep sea anoxic brine of Red Sea and found Siphoviridae and Myoviridae to be dominant viral family members yet appearing distinct from sample to sample. We found that Siphoviridae (32%) in our analysis was the most dominant viral family indicating that perhaps high salt concentration does not put a selective pressure on viral communities. However, these results should be further confirmed with comprehensive analysis involving time series metagenomics of samples collected from different salinity gradients. As mentioned in the results section, we found lambda-like, T4-like, and Cellulophaga phages as some of the most abundant genera among various phage families. In a metagenomic study by Ray et al. (2012) at the Arctic Mid-Ocean Ridge, high abundance of lambda-like phage sequences in both hydrothermal plume and surrounding seawater was also found showing significant ecological role of this phage.
genus in various ecosystems. Furthermore, T4-like phages were previously identified in significant number of sequences in metagenomic study on two freshwater viral communities (Roux et al., 2012). This suggests that viral communities in the hypersaline ecosystem studied in our research does not appear genetically distinct from other aquatic ecosystems. In addition, a previous study conducted with Holmfeldt et al. (2013) revealed that Cellulophaga phage genus was widespread and ubiquitous fraction among marine viral diversity.

For the COG analysis, category A is attributed to RNA processing and was measured in negligible match hits (10 and 15 match hits for bacterial and viral genes, respectively). As chromatin forms chromosomes within the nucleus of eukaryotic cells, category B belongs only to eukaryotes and was detected in minor match hits (8 and 5 match his in bacterial and viral genes, respectively) in our samples. Based on Fisher's exact test, there were significant differences in the functional gene profiles between COG gene categories \( p < 0.05 \). There was an over-representation of genes classified into the “replication, recombination, and repair” category (category L) with 4,341 and 5,022 match hits for bacterial and viral genes, respectively. The most frequent gene among all bacterial predicted genes (occurrence = 416), and the second most frequent gene in the viral predicted genes (occurrence = 354) was related to replicative DNA helicase (COG0305), an enzyme that participates in initiation and elongation during chromosome replication with unwinding DNA and exhibiting DNA-dependent ATPase activity. In addition, bacterial (314 match hits) and viral (340 match hits) metagenomes showed a high representation of the DNA modification methylase (COG0863) gene, which is a part of the restriction-modification systems and responsible for producing a species-characteristic methylation pattern that can be used to protect the bacteria from foreign DNA, such as that borne by bacteriophages.

The COG analysis allowed summarizing the functional gene data and showing the differences in functional trends of bacteria and phage samples. In addition to COG analysis, the predicted genes were also categorized into SEED Subsystems through MG-RAST pipeline. The common proteins in bacterial contigs which were also found in phage contigs were extracted (Table S3). The presence of these enzymes in the bacteria and phage contigs suggests that phage-mediated gene transfer influences environmental processes.

CONCLUSIONS

The Great Salt Lake is one of the most unique hypersaline environments on the planet and has a high diversity in bacteria and bacteriophage species, and many of the bacteria are involved in various nutrient cycles. Metagenomics analyses of bacteria and phage sequences revealed the relationship between the phage and their prokaryotic hosts through their GC content, contig coverage, scaffold length, and tetranucleotide frequency. Furthermore, the presence of prophages and their interaction with the prokaryotic host showed the susceptibility of Great Salt Lake environment to possible prophage induction under different environmental stress factors. In addition, gene prediction in the bacteria and phage contigs showed shared functional genes in the samples, which suggests that gene transfer between bacteria that is mediated by phage transduction and phage-like gene transfer agents (Lang and Beatty, 2007; Stanton, 2007). Lastly, identification of CRISPRs in the bacterial contigs confirms prevalence of previous infections among the prokaryotic community by phages similar to those in marine environments. The phage-host interactions in hypersaline ecosystems are complex and our study gives an enhanced glimpse of phage and prokaryote abundance and diversity as well as their interactions in the hypersaline environment of Great Salt Lake. We also gained interesting information on prophages and how they can affect the prokaryote diversity and nutrient cycling, which are valuable pieces of information for environmental microbiology and ecology.

AUTHOR CONTRIBUTIONS

AM and AB designed and performed laboratory analyses, analyzed data, and contributed to write the manuscript. FC, BD, and SC helped in analyzing and interpreting the data and contributed to writing the manuscript. RG is the primary investigator of this research and contributed in designing the experiments and final approval of the version to be published.

ACKNOWLEDGMENTS

This study was supported by US NSF grant # 1510255, Netherlands Organization for Scientific Research (NWO) Vidi grant 864.14.004 and CAPES/BRASIL. The views and rationale presented in this manuscript are those of the authors and do not necessarily represent the funding agency. We greatly appreciate the help provided by Mr. Ryan Rowland from United States Geological Survey (USGS), Salt Lake City, Utah in collecting sediment samples from Great Salt Lake. We also appreciate the time reviewers have spent on this manuscript to provide their useful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.00352/full#supplementary-material

REFERENCES


Mohaghegh Motlagh et al. Metagenomics of Phage-Host Interactions in Hypersaline Sediments


