RESEARCH ARTICLE

Linking environmental prokaryotic viruses and their host through CRISPRs

Laura Sanguino, Laure Franqueville, Timothy M. Vogel and Catherine Larose*

Environmental Microbial Genomics, Laboratoire Ampère, CNRS UMR 5005, Ecole Centrale de Lyon, Université de Lyon, 36 Avenue Guy de Collongue, 69134 Ecully, France

*Corresponding author: Environmental Microbial Genomics, Laboratoire Ampère, CNRS UMR 5005, Ecole Centrale de Lyon, Université de Lyon, 36 Avenue Guy de Collongue, 69134 Ecully, France. Tel: +33-4-72-18-62-37; Fax +33-4-78-43-37-17; E-mail: catherine.larose@ec-lyon.fr

One sentence summary: We developed a method to recover data on virus-host interactions from environmental samples using CRISPR sequences. We applied this workflow to Arctic ice and soil metagenomes.

Editor: Julie Olson

ABSTRACT

The ecological pressure that viruses place on microbial communities is not only based on predation, but also on gene transfer. In order to determine the potential impact of viruses and transduction, we need a better understanding of the dynamics of interactions between viruses and their hosts in the environment. Data on environmental viruses are scarce, and methods for tracking their interactions with prokaryotes are needed. Clustered regularly interspaced short palindromic repeats (CRISPRs), which contain viral sequences in bacterial genomes, might help document the history of virus-host interactions in the environment. In this study, a bioinformatics network linking viruses and their hosts using CRISPR sequences obtained from metagenomic data was developed and applied to metagenomes from Arctic glacial ice and soil. The application of our network approach showed that putative interactions were more commonly detected in the ice samples than the soil which would be consistent with the ice viral–bacterial interactions being more dynamic than those in soil. Further analysis of the viral sequences in the CRISPRs indicated that *Ralstonia* phages might be agents of transduction in the Arctic glacial ice.

Keywords: virus-host interactions; CRISPRs; transduction; glacial ice

INTRODUCTION

Viruses outnumber prokaryotes and are the most abundant biological entities in a range of ecosystems (Fuhrman 1999; Williamson *et al.* 2013) and possibly in the entire biosphere. Through infection and microbial lysis, they can regulate microbial populations as well as generate and maintain microbial diversity (Thingstad 2000; Rodriguez-Valera *et al.* 2009), which in turn has an impact on food webs and biogeochemical cycles (Weinbauer 2004; Suttle 2007). Furthermore, viruses can be agents of horizontal gene transfer through transduction by packaging microbial genetic material into their capsid during

the lytic cycle and transferring it to another host (Canchaya *et al.* 2003). The rate of gene transfer by transduction in oceans has been calculated to be 10^{24} genes per year (Jiang and Paul 1998; Rohwer and Thurber 2009), which was concluded to be an important mechanism for microbial adaptation. Examples of specific viral transduction events have also been reported: marine viruses carry a range of carbon metabolism genes (Lindell *et al.* 2005; Hurwitz, Hallam and Sullivan 2013); transduction of toxin genes has been seen to transform bacteria into pathogens, thus extending the niches of their hosts (Waldor and Mekalanos 1996). Viruses might be drivers of microbial adaptation to

environmental conditions by enabling the acquisition of whole segments of genomic information that can recombine or integrate into the host genome.

In order to evaluate the impact of viruses on microbial adaptation and possibly evolution, we need further information about environmental virus-host interactions. However, at least 50% of the sequences retrieved from publically available viral metagenomes have no similarities with specific viral sequences in databases (Rosario and Breitbart 2011; Ray *et al.* 2012), rendering the study of environmental viruses through high-throughput methods challenging. Moreover, relative host-species specificities in uncultured bacteria and viruses remain unknown. Clustered regularly interspaced short palindromic repeats (CRISPRs), which play a role in the defense of prokaryotes against viruses, have been suggested as a potential tool to identify virus-host interactions in complex environmental communities (Anderson, Brazelton and Baross 2011). The system is heritable and widespread (Fineran and Charpentier 2012) and nearly all archaea and about half of the bacteria have been found to have CRISPR-Cas systems (Jansen *et al.* 2002; Terns and Terns 2011). CRISPRs are made of a series of direct repeats (DRs) of microbial origin interspaced by short sequences called spacers of viral origin (among others) (Mojica *et al.* 2005). After an infection, short sequences from the virus (proto-spacers) are inserted into the CRISPR. These short specific sequences enable the host to recognize the virus and protect itself against future infections (Barrangou *et al.* 2007) by silencing the foreign DNA through the Cas enzymatic machinery (Terns and Terns 2011). In order to test whether we could link viruses and their hosts through CRISPRs, we used Arctic glacier ice and soil as model ecosystems. Studies about viruses in polar environments support the idea of high viral diversity (López-Bueno *et al.* 2009), high infection rates (Säwström *et al.* 2007; Bellas *et al.* 2013) and broad host ranges (Anesio *et al.* 2007; Colangelo-Lillis and Deming 2013) that would enable them to infect different bacteria within the same environment and possibly bacteria from different environments (Sano *et al.* 2004; Anesio *et al.* 2007; Bellas and Anesio 2013). These characteristics become especially relevant when studying viral transduction and explain why this environment is an interesting model ecosystem for this study. Furthermore, we hypothesize that an environment with a high diversity of viruses with broad host ranges would harbor a higher number of unique CRISPR sequences. This hypothesis is based on the idea that to protect itself from a viral infection, a microorganism has to have a spacer for each different virus that could (or did) infect it (Barrangou *et al.* 2007). If the diversity of viruses is high and they can infect several different microbes, the number of spacer combinations in CRISPR arrays would also be high, giving place to an increase in the number of different CRISPRs. Therefore, we hypothesize that the relative number of unique CRISPRs found in a metagenome could be an indicator for viral diversity and for the dynamism of the interactions with their hosts.

We developed a method to construct infection networks using spacers and DRs outsourced from CRISPR sequences from two different microbial and viral metagenomes. We applied this workflow to Arctic ice and soil metagenomes. This workflow identified putative virus-host interactions and provided the basis for comparing soil and ice bacterial/viral interaction dynamics. Gene transfer events were also evaluated in an effort to further define the nature of these interactions in Arctic ice.

MATERIAL AND METHODS

Study site

Ice samples were recovered from Kongsvegen glacier in Svalbard (N78°45′19.8″, E13°20′11″), Norway, in the spring of 2007. A 15 m ice core was extracted from the accumulation area of the glacier, at an altitude of 670 m over sea level, using a manual drill (PICO corer) without any drilling fluids. The ice was stored in clean, sterile bags and frozen (−20°C) until further use. The core was composed of ice and firn layers; only ice layers were used in this study. On the other hand, soil was collected at the outskirts of Ny-Alesund (N78°45′20.16″, E13°20′11.4″) and was also stored frozen (−20°C). The soil was covered by snow and had a lichen crust over it. Samples were recovered from the first 5–10 cm and were found to be very rich in clay and organic matter.

Filtration, DNA extraction and sequencing

Microbial and viral DNA from ice

A total of 9 l of ice was melted at room temperature and filtrated simultaneously through 10 μm 47 mm (Millipore) filters and through 0.2 μm 47 mm (Millipore) to recover the bacterial fraction. Filtrate was then filtered through 0.02 μm 25 mm (Anodisc, Anopore Inorganic Membranes, Whatman) to recover viral fraction. The 0.02 μm filters were left overnight in PBS (without Ca or Mg, autoclaved and filtered through 0.02 μm) and then put in the shaker for 2 h before discarding the filter and proceeding with the extraction protocol. To assess the presence of viral particles and/or microbial cells, aliquots of the samples were filtered through 0.02 μm filters, stained with SYBER Gold (Life Technologies) and examined through epifluorescence microscopy as described by Thurber *et al.* (2009).

Viral DNA was extracted following the protocol from Thurber *et al.* (2009), yielding 0.502 ng μl^{-1} of DNA based on Qubit DNA HS fluorometric quantitation (Life Technologies). Multiple displacement amplification (MDA) was performed for viral DNA (illustra GenomiPhi HY DNA Amplification Kit, GE Healthcare Life Sciences, NJ, USA) following manufacturer instructions in order to have enough DNA for sequencing. DNA was then cleaned with GeneClean Turbo Kit (MP Biomedicals, France). Microbial DNA was extracted from two 10 μm and two 0.2 μm filters as described by (Larose *et al.* (2010)), DNA concentrations were of the order of 2 ng μl^{-1} and samples were stored at −20°C until further use. DNA from viral and microbial extractions was sent for sequencing using GS FLX System (Roche-454 Life Science, Beckman Coulter Inc., MA, USA).

Microbial and viral DNA from soil

A total of 900 g of soil was used for DNA extraction, following a modified protocol from Thurber *et al.* (2009). Briefly, 20 ml of PBS (without Ca or Mg in order to decrease the adherence of cells to soil particles, autoclaved and filtered through 0.02 μm filters) was added to every 50 g of soil and then blended with a Waring blender. The mixture was then centrifuged for 10 min at 2500 g and the supernatant was retrieved with sterile syringes. The supernatant was filtered through 0.45 μm 47 mm filters (Millipore) and then through 0.2 μm 47 mm filters. Microbial DNA was recovered from the filters while viral DNA was recovered from the filtrate. Microbial extraction of soil was done along with microbial extraction of ice as described by Larose *et al.* (2010). The extraction yielded 28 ng μl^{-1} of microbial DNA. Examination of the viral concentrate through epifluorescence microscopy was done
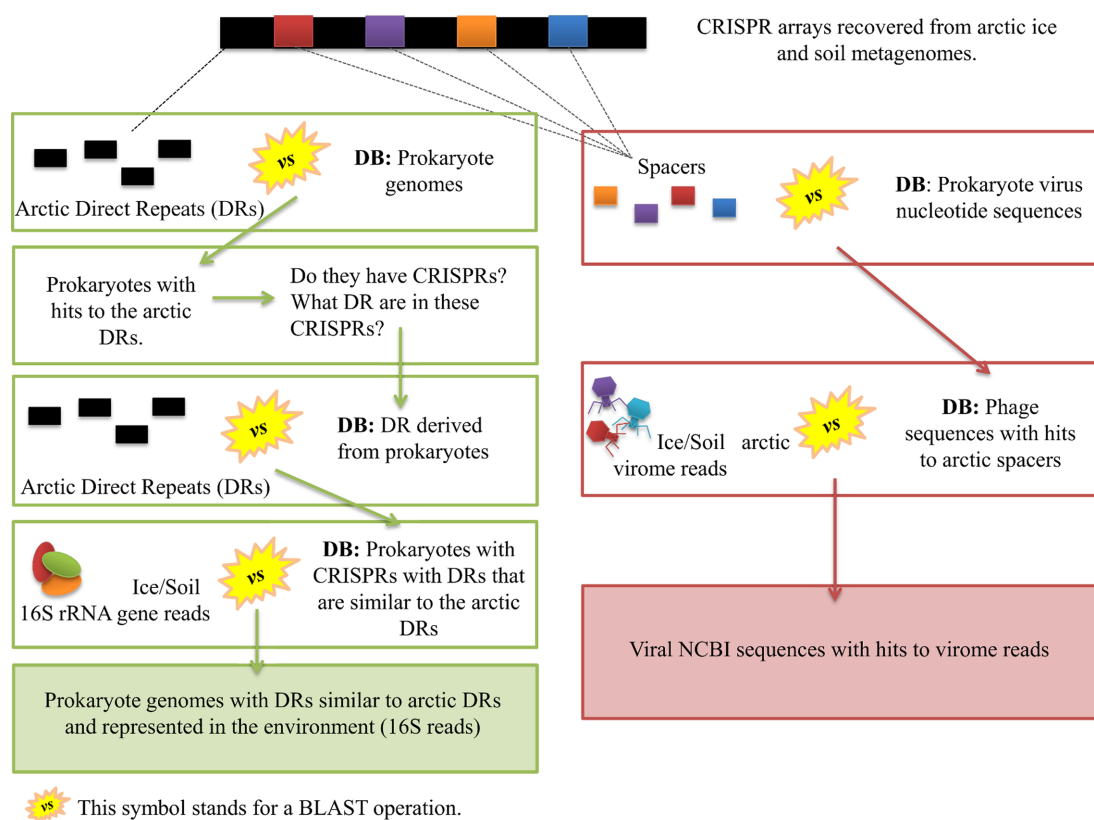
Figure 1. Diagram explaining the workflow used to identify the putative origin of the DRs (left) and the putative origin of spacers (right). This information was used to produce the networks linking phages to their hosts (Fig. 2). DB stands for database.

as described above. Viral DNA was extracted following the protocol from Thurber *et al.* (2009), DNA concentrations were 3.83 ng μl^{-1}, and the samples were stored at -20°C until further use.

16S rRNA sequencing from ice and soil

The 454 Pyrotag sequencing was done for regions V4 to V6 from the 16S rRNA genes from soil and ice microbial DNA extracts. Barcoded amplicons were prepared using MID-primers; each 50 μl PCR reaction contained 37 μl water, 5 μl Titanium *Taq* PCR Buffer, 1 μl of each MID-primer pair, 1 μl of dNTP Mix, 1 μl Titanium *Taq* DNA Polymerase (Clontech) and 4 μl (10 ng) of template DNA. The PCR cycling conditions were: 1 min at 95°C for denaturation followed by 30 cycles of 30 s denaturation at 95°C, 3 min annealing and elongation at 68°C and one cycle of 3 min at 68°C. The amplified DNA was size selected by excising the appropriate gel bands and purified using Illustra GFX PCR DNA and gel band purification kit. The DNA was quantified using Qubit DNA HS assay (Life Technologies). Amplicons were pooled in equimolar concentrations. The amplicon libraries were sequenced by Beckman Coulter Inc. (MA, USA) using 454 platforms. Sequencing data were analyzed following QIIME pipeline (Caporaso *et al.* 2011) for *de novo* OTU picking and diversity analysis using 454 data. Briefly, OTUs were picked based on sequence similarity and taxonomy was assigned through uclust consensus taxonomy classifier.

CRISPR infection network development

CRISPRs were browsed in the ice and soil metagenomes using metaCRT (Rho *et al.* 2012). It was seen that the assembly of the

metagenomes into contigs significantly reduced the number of CRISPRs found; thus, only the individual reads were used for this purpose. The resulting CRISPR arrays were decomposed into a DR file and a spacer file. The workflow followed is summarized in Fig. 1. Ten CRISPRs were found in the ice metagenome while one CRISPR was found in the soil metagenome (Fig. 2). There were no repeated CRISPR sequences.

DRs have been seen to be conserved among taxa (Mojica, Ferrer and Juez 1995; Kunin, Sorek and Hugenholtz 2007), so they could serve to identify the organisms to whom the different CRISPR arrays belonged to by blasting the DR against an archaeal and bacterial genome database. However, there is also the possibility that a DR is shared among different bacteria. The database we used was a GenBank database (5193 full genomes), downloaded from their site in May 2014. The organisms retrieved through the blast of the DRs against the database were considered candidate sources of the CRISPRs (default stringency parameters: maximum *e*-value 10). We then verified if these candidate bacteria had CRISPR arrays in their genomes and made a database of the DR found in these CRISPRs. The DRs we had retrieved from our metagenomic CRISPRs were blasted against this database (default stringency parameters: maximum *e*-value 10). This left us with a list of prokaryotic genomes with CRISPRs whose DRs were similar to the DR we had outsourced from our metagenomic CRISPRs. We then checked to see if these prokaryotic genomes were represented in our 16S rRNA gene libraries by doing a blast of the 16S rRNA gene sequences against the prokaryotic genomes we had left at this point as putative originators of the metagenomics CRISPRs (97% identity over a minimum of 200 bp). While we had no hits for the soil, there were 47 genomes represented in our 16S rRNA gene library that had
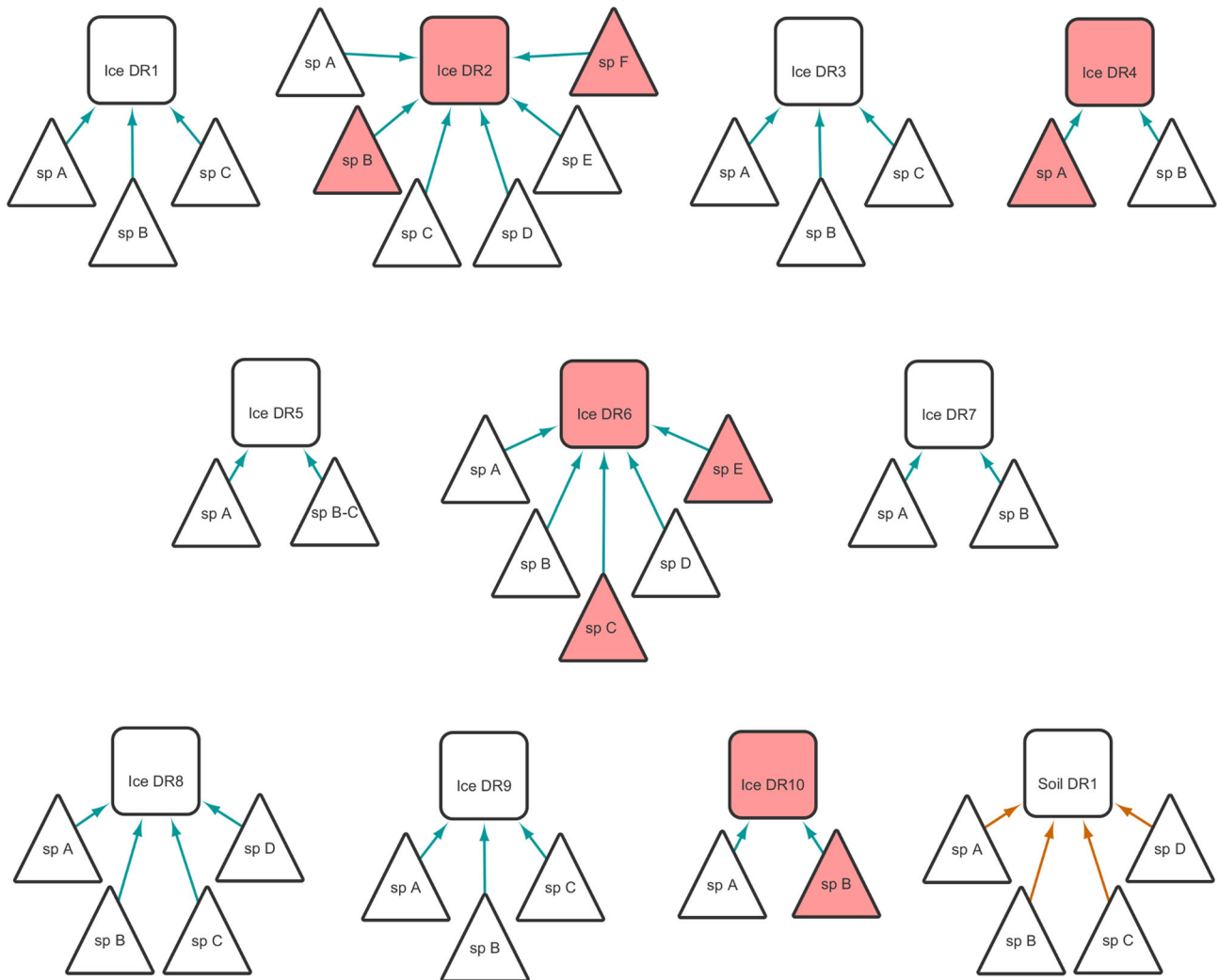
Figure 2. Set of networks representing the CRISPRs found in ice (blue arrows) and soil (brown arrows). Rectangles represent the DRs and triangles represent the spacers. Where identification of the microbial originator of the DR and the viral originator of the spacer was possible, nodes are shaded in red.

CRISPRs with DR similar to the DR of five different CRISPRs from the ice metagenomes. These 47 genomes belonged to 11 different prokaryotic species spread through six genera.

The spacers are sequences that can belong to viruses, and therefore, we used them to putatively identify potential viral infectors of microbial cells. Given that these are small viral DNA fragments, the sequences might be found in more than one virus. Spacers outsourced from the CRISPRs from the ice were blasted against a prokaryotic viruses' GenBank database sourced from NCBI in May 2014 (12.954 nucleotide sequences). Given that we were unable to identify the microbial source of the DR for the soil samples, only the spacers from the ice samples were analyzed. The viruses recovered (default stringency parameters: maximum e-value 10) were considered to be putative sources of the spacers. In a similar way as for the DR, we checked if these viruses were represented in our metagenomes. Since there is no analogous sequence to the 16S rRNA gene for viruses, we decided to do a blast of the viral metagenomes against the putative sources (maximum e-value 9×10^{-10} to account for the small size of the database). A total of 15 viruses were represented in our data and were considered as potential sources of the spacers from the ice samples, outsourced from four different CRISPRs. The 15 viruses belonged to one of these

seven types of viruses: *Mycobacterium* phage, *Cellulophaga* phage, Cyanophage, *Synechococcus* phage, *Pseudomonas* phage, environmental Halophage and *Geobacillus* virus.

Finally, the infection network was built using the software Cytoscape (Saito *et al.* 2012). The four CRISPRs (I to IV, Fig. 3) for which we had identified both the origin of the spacers and of the DR were represented in networks (Fig. 3).

Searching for transduction events

Transduction can take place through two different mechanisms. Generalized transduction is when a virus encapsidates pieces of microbial DNA instead of its own genome. On the other hand, when a temperate virus takes an adjacent piece of microbial DNA along with its viral genome, it is called specialized transduction. Thus, sequences sharing microbial and viral information could stand as a hallmark for specialized transduction; therefore, we looked for such hallmarks in our viral contigs. Viral 454 reads were assembled into contigs using the software provided by Roche Diagnostics Corporation, GS De Novo Assembler. Sequence redundancy had been previously reduced using CD-HIT-454 (Fu *et al.* 2012). Assembly of the sequences from the ice yielded 689 contigs, with a mean length of 1423 bp. In the
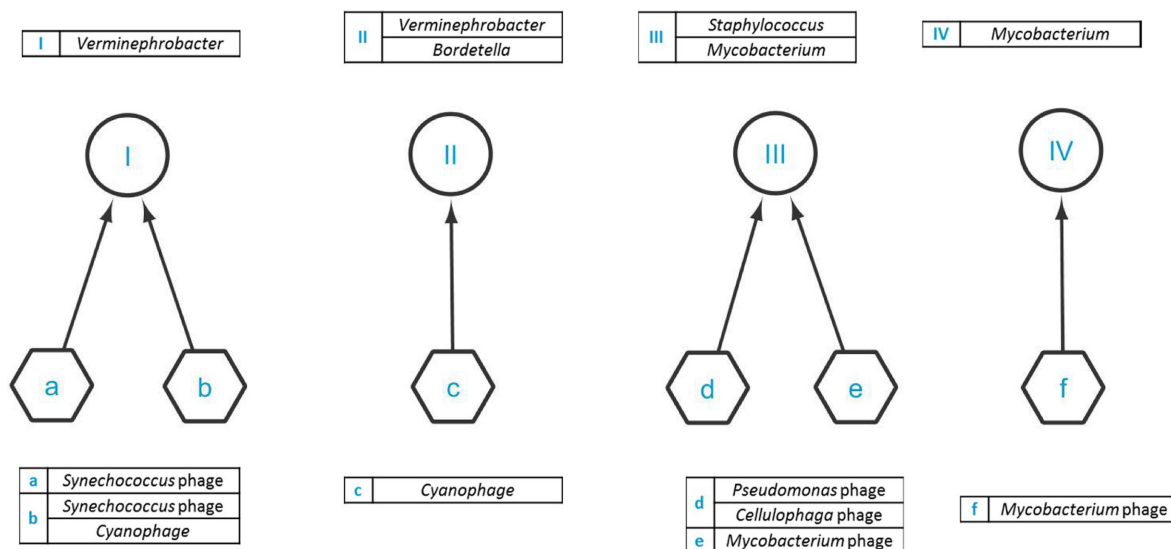
Figure 3. Networks with both the origin of the DR and at least one related spacer were putatively identified. Above and below each network, legends show the identified prokaryotes (roman numerals) and the identified viruses (letters).

case of viral soil sequences, assembly yielded 2076 contigs with a mean length of 711 bp. GenBank protein sequences for mercury resistance, cold shock and efflux pumps genes were downloaded from NCBI in May 2014. Each of these sets of sequences was blasted against the ice and the soil viral contigs. Contigs with hits to the protein sequences (maximum blast *e*-value 9×10^{-10}) were taken as contigs with microbial DNA within them. In order to test if these contigs also had viral DNA within them, we blasted them against the prokaryotic virus database. The contigs with hits to the prokaryotic virus database (maximum blast *e*-value 9×10^{-10}) were taken as possible hallmarks of transduction since they shared both microbial and viral information. The viral part of the contig enabled us to identify the possible transduction agents.

RESULTS

CRISPR infection network

The aim of this study was to develop a method to use CRISPRs to decipher virus-host interactions in complex environments. Our CRISPR-based identification workflow (explained in detail in the section 'Material and Methods') was used to determine the possible bacterial and viral origin of DRs and spacers, respectively. CRISPRs extracted with the help of the software metaCRT were then analyzed through our bioinformatics pipeline. The CRISPRs from the ice (blue arrows) and soil (brown arrows) microbial metagenomes were used to define possible bacterial hosts (rectangles representing the DRs from the bacteria) and the infecting viruses (triangles) (Fig. 2). While this figure depicts the interactions that could be taking place in the environment, this network is a simplified diagram of Arctic virus-host interactions as each DR could be common to several bacteria and every spacer could be common to different viruses. So while the network provides the possible connection between DRs and spacers from ice and soil, the number of putative bacteria and viruses involved is unknown. For further analysis, the networks (shaded in red in Fig. 2) with a DR that was found in the bacterial genome database (5193 full genomes) and with a spacer also found in the viral

database were recreated with their possible bacterial and viral identities (Fig. 3). Thus, circles stand for the microbial host and hexagons stand for the microbial viruses (Fig. 3). The roman numerals within the circles relate to the legends above the network, where the putative bacteria identified with the DRs are listed. The lowercase letters within the hexagons relate to the legends below the network, where the putative viruses identified with the spacers are listed. An added stringency to the data presented in Fig. 3 was the requirement that the putative bacteria from the 5193 full genomes had a matching 16S rRNA gene sequence in the pyrotags sequenced from our ice and soil samples. The networks in Fig. 3 belong exclusively to ice virus-host interactions as none of the putative soil bacteria matched the 16S rRNA sequences in our pyrotag data from the Arctic soil.

The genus *Verminephrobacter* was identified as a possible microbial host in I and II (Fig. 3). In the case of II, the microbial host could also be *Bordetella*. However, viruses infecting I and II were identified to be *Cyanobacterial* phages. While three *Bordetella* phage genomes were present in the prokaryotic viruses' database, there were no known *Verminephrobacter* phages. Thus, viruses a, b and c might be *Verminephrobacter* phages, although their closest match in the viral database are *Cyanobacteria* phages. Microbial host III was identified as a member of either *Staphylococcus* or *Mycobacterium*. Microbial host IV was also identified as a *Mycobacterium*; congruently viruses e and f were identified as *Mycobacterium* phages. Phage d was assigned to either *Cellulophaga* phage or *Pseudomonas* phage. Further interrogation of the ice metagenomes through MG-RAST (Meyer *et al.* 2008) confirmed the assignment of reads to all identified hosts. Likewise, viruses identified from the spacers were represented in the viral ice metagenomes based on Metavir's (Roux *et al.* 2011) contig and/or read assignments, supporting data output from the CRISPR network system.

Potential transduction agents in ice and soil

In specialized transduction, the virus encapsidates its own genome with part of the host microbial genome (Morse,

Table 1. Hybrid contigs with both microbial and viral sequences, taken as hallmarks of specialized transduction. **a.** Contigs with cold shock protein sequences and their viral assignment. **b.** Contigs with *mer* genes' protein sequences and their viral assignment. **c.** Contigs with efflux pumps' protein sequences and their viral assignment.

Contig	GenBank accession	Definition
a. COLD SHOCK PROTEINS		
contig00022	AB259123.2	*Ralstonia* phage
contig00022	KF887906.1	*Ralstonia* phage
contig00022	AB434711.1	*Ralstonia* phage
contig00006	AB276040.1	*Ralstonia* phage
contig00500	GU396103.1	*Aeromonas* phage
contig00157	GU396103.1	*Aeromonas* phage
b. *mer* GENE PROTEINS		
contig00022	AB259123.2	*Ralstonia* phage
contig00022	KF887906.1	*Ralstonia* phage
contig00022	AB434711.1	*Ralstonia* phage
c. EFFLUX PUMP PROTEINS		
contig00006	AB276040.1	*Ralstonia* phage
contig00007	JN662425.1	*Burkholderia* phage
contig00007	FJ937737.2	*Burkholderia* phage
contig00007	JX104231.1	*Burkholderia* phage
contig00022	AB259123.2	*Ralstonia* phage
contig00022	KF887906.1	*Ralstonia* phage
contig00022	AB434711.1	*Ralstonia* phage
contig00035	AB828698.1	*Ralstonia* phage
contig00035	GU911303.1	*Burkholderia* phage
contig00035	J01735.1	Bacteriophage lambda
contig00035	AF234172.1	*Enterobacteria* phage
contig00035	AF234173.1	*Enterobacteria* phage
contig00157	GU396103.1	*Aeromonas* phage

Lederberg and Lederberg 1956; Frost *et al.* 2005). If these events had taken place in the microbial communities we sampled in the Arctic, we would expect to find viral genomes with fragments of microbial DNA. In order to look for traces of specialized transduction events in our viral metagenomes, we selected a set of genes that we thought could be of adaptive interest to the Arctic microbial community. As the Arctic ice and soil are subjected to low temperatures and mercury from the snowpack (Larose *et al.* 2013), we selected cold shock genes and mercury resistance genes as traits that might be an asset for microorganisms living in the Arctic. Additionally, we selected a set of genes, such as efflux pumps genes, that might provide an adaptive advantage regardless of the ecosystem. We browsed our assembled viral metagenomes for hybrid contigs and subsequently identified the transduction agent through the viral part of the contig. The viruses that could be potential transducing agents for the three selected genes are presented in Table 1 a, b and c. No soil contigs contained both a viral and microbial marker based on our stringency parameters (see the section 'Material and Methods'). While in the case of the ice metagenome, there were six contigs that had putative viral and microbial markers matching those in the microbial and the viral databases. *Ralstonia* phage was identified as a transducing agent for all three microbial proteins studied here (Table 1 a, b and c).

DISCUSSION

The impact of prokaryotic viruses on the ecology and evolution of microbial communities is tightly linked to their infectivity potential and their host range (Weinbauer and Rassoulzadegan 2003; Koskella and Meaden 2013). In order to address their impact, we need to link viruses to their microbial host(s) in the environment. The aim of this paper was to demonstrate the potential of using CRISPRs as a tool to link viruses with their hosts in environmental samples. Given that this demonstration is a proof of concept, we were quite stringent in our workflow in order to clearly show both the application to unknown viruses and bacteria as well as with bacterial and viral sequences of putative origin. Since CRISPRs contain sequences derived from bacteria and archaea (in the DRs) (Jansen *et al.* 2002) and spacers from viruses, we initially constructed the networks between DRs and spacers based on identified CRISPRs from the different metagenome datasets that we sequenced from Arctic glacial ice and soil (Fig. 2). With similar quantity of metagenomic sequences from ice and soil (42 078 reads for soil and 47 626 reads for ice), we found more CRISPRs in ice than in soil and all CRISPR sequences found were unique. To take into account the possible influence of MDA on the number of CRISPRs found, a control (soil DNA subject to MDA) was evaluated along with the other samples. Two CRISPRs were found in the MDA control compared to the one CRISPR found in the soil (not subjected to MDA), so the MDA did not significantly change the number of CRISPRs detected. Further work on the environmental aspects of CRISPRs needs to be carried out to evaluate the different CRISPR abundances in different environments and its consequences. However, we hypothesize that a higher number of unique CRISPR sequences in a metagenome might represent higher viral diversity and interactions with their hosts through the infection of different microorganisms. Thus, the relatively more abundant CRISPRs in ice might be an indication of the high viral diversity, infectivity rates and broad host ranges as suggested in the literature (Anesio *et al.* 2007; Säwström *et al.* 2007; López-Bueno *et al.* 2009). When we restrained the possible bacterial and archaeal hosts to those that had their genomes fully sequenced and reported in the NCBI database and to the viruses represented in the NCBI viral sequences, the number of networks was significantly reduced (Fig. 3). These prokaryotic genomes were used to identify potential originators of the DRs from the CRISPRs found in the ice and soil metagenomes, but only if similar 16S rRNA genes were also identified in the pyrotag libraries obtained for each environmental sample. Viral databases are, for the moment, very scant, making it difficult to identify viruses through DNA sequences such as spacers, especially in the environment. However, despite working with a relatively unexplored ecosystem whose inhabitants may not be well represented by the fully sequenced genomes in the database, we were able to describe several virus-host interactions taking place in the Arctic glacial ice through the analysis of CRISPRs obtained from metagenomic data.

With the requirement that the DRs be found in fully sequenced bacterial genomes in the NCBI database and that our 16S rRNA gene pyrotags from the Arctic match the 16S rRNA sequences from those fully sequenced genomes, our CRISPR network (10 CRISPRs from the ice and 1 from the soil) was reduced to four CRISPRs in ice and none in soil (Fig. 3). The low number of CRISPRs found in soil might be due to a limited depth of coverage in the more diverse soil metagenome dataset. Therefore, ice appears to be a good model ecosystem for this initial CRISPR application study. The analysis of ice CRISPRs shed some light on the communities inhabiting this environment and their possible viral relationships. *Verminephrobacter* (possible host I and II in Fig. 3) has been described as a nephridia symbiont, suggesting the presence of annelids in the ice. No functional data

for annelids were present in our metagenome dataset based on an analysis with MG-RAST. However, given that samples were pre-filtered to remove debris, it is possible that we lost these in the process. The presence of ice worms has been confirmed in a range of glaciers (Hartzell et al. 2005), and representatives of *Verminephrobacter* were also found in Lake Vostok ice (Shtarkman et al. 2013). Recently, a study described the presence of bacterial microbiota in ice worms (Murakami et al. 2015). *Bordetella* (possible host II) was also found to be present in ice cores from Pony Lake, Antarctica (Foreman et al. 2011). Although *Staphylococcus* and *Mycobacterium* (possible host III and possible host III and IV, respectively) are genera often associated with humans, they are ubiquitous genera that have been found inhabiting a wide range of environments (Kloos 1980; Falkinham 2002), including glacial ice (Christner et al. 2000; Miteva, Sheridan and Brenchley 2004; Xiang et al. 2005). When looking at the taxonomic assignment of the 16S rRNA gene sequences, only *Mycobacterium* was present at the genus level (0.14% of assigned reads). No other bacteria identified through our network were represented in the taxonomic classification of the 16S rRNA gene sequences, which could point to a low abundance of these bacteria in their environment. It could also mean that these bacteria are the closest members present in the database, but that the bacteria actually inhabiting the ice do not have representatives in the 16S rRNA gene database, thus limiting the taxonomic assignment. With respect to the viruses identified in our network, all of them, with the exception of the *Cellulophaga* phage, were described by Cottrell and Kirchman (2012) in their analysis of viral genes present in Arctic bacteria. These data support the assumption that these viruses are infecting Arctic bacteria as indicated by our results. In addition, *Cellulophaga* phages seem to be ubiquitous in aquatic environments based on the study by Holmfeldt et al. (2013).

Our interest in linking phages and their host in the environment originated from the hypothesis that transduction has a significant role in microbial adaptation. We, therefore, tried to identify molecular tracers of transduction in our viral metagenomes. We found hybrid contigs in our assembled virome that had both viral and microbial sequences in the same contig. From this search for transducing agents, *Ralstonia* phages were frequent candidates in our Arctic viromes. No hybrid contig was found in the soil virome while six hybrid contigs were found in the ice virome. Given that the overall number of viral ice contigs was roughly a third of the contigs from the soil, transduction in ice might be more predominant than in the soil, where other mechanisms of gene transfer might compete with transduction. Based on this result, *Aeromonas* and *Ralstonia* phages in Arctic glacial ice appear to be carrying pieces of various microbial genes that could be (or have been) transferred to their host(s). 16S rRNA genes assigned to *Aeromonas* were found in viral particles issued from wastewater (Sander and Schmieger 2001). In addition, individual *Ralstonia* phages appear to be capable of infecting several different strains of *Ralstonia solanacearum* (Kawasaki et al. 2009) and possibly spreading important environmental genes throughout the population. Moreover, *Cupriavidus metallidurans*, formerly *Ralstonia metallidurans* (Vandamme and Coenye 2004), is a bacteria known to resist a range of different metals including mercury by carrying the genes responsible for this resistance in a *mer* operon (von Rozycki and Nies 2009). Given that we could find *Cupriavidus metallidurans* in our MG-RAST functional assignment data (Meyer et al. 2008), *Ralstonia* phages found in this study as transduction agents of mercury resistance genes might have *Cupriavidus metallidurans* as a host.

The stringent nature of the data interpretation here has eliminated possible bacterial/viral interactions when both the bacteria and the virus were unknown or the bacterial presence was not confirmed by 16S rRNA sequencing. Therefore, this reduced output results in part from the rather limited databases available. The increase in viruses included in viral databases as well as bacterial genomes in genomic databases will improve the effectiveness of CRISPRs as a tool to link known viruses to their known hosts. CRISPRs are also quite abundant in archaea, but the number of fully sequenced archaeal genomes is considerably smaller than that for bacteria. In a similar way, deeper microbial sequencing could also increase the number of CRISPR sequences found in metagenomes and make the identification of the microbial originator of the CRISPR a more likely occurrence in environments with high microbial diversity. On the other hand, our CRISPR-based network discovered viral/microbial interactions among unknown viruses and bacteria/archaea. The further identification of these microorganisms would help clarify the extent of these interactions.

The results described here are consistent with Arctic ice being an ecosystem with virus-host interactions and that these interactions seem to be relatively (on a per mass of DNA basis) more productive than the ones taking place in nearby Arctic soil. In addition, the transduction of the genes studied here might be more prevalent in ice than in soil as a consequence of more dynamic interactions. This would be consistent with previous work that suggested virus-host interactions in cold environments could be highly dynamic (Wells and Deming 2006; Anesio and Bellas 2011). Our study has developed a CRISPR-based tool to link viruses with their host(s) in the environment through metagenomic data. In doing so we were able to provide a new perspective and gather information on virus-host relationships. This new perspective could help us further understand and compare viral ecology in different ecosystems, giving us a more complete view of microbial communities, their interactions, and consequently, their influence on global processes.

REFERENCES

Anderson RE, Brazelton WJ, Baross JA. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol Ecol* 2011;**77**:120–33.

Anesio AM, Bellas CM. Are low temperature habitats hot spots of microbial evolution driven by viruses? *Trends Microbiol* 2011;**19**:52–7.

Anesio AM, Mindl B, Laybourn-Parry J, *et al.* Viral dynamics in cryoconite holes on a high Arctic glacier (Svalbard). *J Geophys Res* 2007;**112**:G04S31.

Barrangou R, Fremaux C, Deveau H, *et al*. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;**315**:1709–12.

Bellas CM, Anesio AM. High diversity and potential origins of T4-type bacteriophages on the surface of Arctic glaciers. *Extremophiles* 2013;**17**:861–70.

Bellas CM, Anesio AM, Telling J, *et al*. Viral impacts on bacterial communities in Arctic cryoconite. *Environ Res Lett* 2013;**8**:045021.

Canchaya C, Fournous G, Chibani-Chennoufi S, *et al*. Phage as agents of lateral gene transfer. *Curr Opin Microbiol* 2003;**6**: 417–24.

Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2011;**7**:335–6.

Christner BC, Mosley-thompson E, Thompson LG, *et al*. Recovery and identification of viable bacteria immured in glacial ice. *Icarus* 2000;**144**:479–85.

Colangelo-Lillis JR, Deming JW. Genomic analysis of cold-active Colwelliaphage 9A and psychrophilic phage-host interactions. *Extremophiles* 2013;**17**:99–114.

Cottrell M, Kirchman D. Virus genes in Arctic marine bacteria identified by metagenomic analysis. *Aquat Microb Ecol* 2012;**66**:107–16.

Falkinham JOI. Nontuberculous mycobacteria in the environment. *Clin Chest Med* 2002;**23**:529–51.

Fineran PC, Charpentier E. Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology* 2012;**434**:202–9.

Foreman CM, Dieser M, Greenwood M, *et al*. When a habitat freezes solid: microorganisms over-winter within the ice column of a coastal Antarctic lake. *FEMS Microbiol Ecol* 2011;**76**:401–12.

Frost LS, Leplae R, Summers AO, *et al*. Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005;**3**:722–32.

Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.

Fuhrman JA. Marine viruses and their biogeochemical and ecological effects. *Nature* 1999;**399**:541–8.

Hartzell PL, Nghiem JV, Richio KJ, *et al*. Distribution and phylogeny of glacier ice worms (Mesenchytraeus solifugus and Mesenchytraeus solifugus rainierensis). *Can J Zool* 2005;**83**:1206–13.

Holmfeldt K, Solonenko N, Shah M, *et al*. Twelve previously unknown phage genera are ubiquitous in global oceans. *P Natl Acad Sci USA* 2013;**110**:12798–803.

Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol* 2013;**14**:R123.

Jansen R, Embden JD, Gaastra W, *et al*. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 2002;**43**:1565–75.

Jiang SC, Paul JH. Gene transfer by transduction in the marine environment. *Appl Environ Microb* 1998;**64**:2780–87.

Kawasaki T, Shimizu M, Satsuma H, *et al*. Genomic characterization of Ralstonia solanacearum phage phiRSB1, a T7-like wide-host-range phage. *J Bacteriol* 2009;**191**:422–7.

Kloos WE. Natural populations of the genus Staphylococcus. *Annu Rev Microbiol* 1980;**34**:559–92.

Koskella B, Meaden S. Understanding bacteriophage specificity in natural microbial communities. *Viruses* 2013;**5**: 806–23.

Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 2007;**8**:R61.

Larose C, Berger S, Ferrari C, *et al*. Microbial sequences retrieved from environmental samples from seasonal arctic snow and meltwater from Svalbard, Norway. *Extremophiles* 2010;**14**: 205–12.

Larose C, Prestat E, Cecillon S, *et al*. Interactions between snow chemistry, mercury inputs and microbial population dynamics in an Arctic snowpack. *PLoS One* 2013;**8**:e79972.

Lindell D, Jaffe JD, Johnson ZI, *et al*. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 2005;**438**:86–9.

López-Bueno A, Tamames J, Velázquez D, *et al*. High diversity of the viral community from an Antarctic lake. *Science* 2009;**326**:858–61.

Meyer F, Paarmann D, D'Souza M, *et al*. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;**9**:386.

Miteva VI, Sheridan PP, Brenchley JE. Phylogenetic and physiological diversity of microorganisms isolated from a deep Greenland glacier ice core. *Appl Environ Microb* 2004;**70**: 202–13.

Mojica FJM, Díez-Villaseñor C, García-Martínez J, *et al*. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 2005;**60**: 174–82.

Mojica FJM, Ferrer C, Juez G. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloforax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. 1995;**17**:85–93.

Morse ML, Lederberg EM, Lederberg J. Transduction in Escherichia coli k-12. *Genetics* 1956;**41**:142–56.

Murakami T, Segawa T, Bodington D, *et al*. Census of bacterial microbiota associated with the glacier ice worm Mesenchytraeus solifugus. *FEMS Microbiol Ecol* 2015;**91**:1–10.

Ray J, Dondrup M, Modha S, *et al*. Finding a needle in the virus metagenome haystack–micro-metagenome analysis captures a snapshot of the diversity of a bacteriophage armoire. *PLoS One* 2012;**7**:e34238.

Rho M, Wu Y-W, Tang H, *et al*. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet* 2012;**8**:e1002441.

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, *et al*. Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 2009;**7**:828–36.

Rohwer F, Thurber RV. Viruses manipulate the marine environment. *Nature* 2009;**459**:207–12.

Rosario K, Breitbart M. Exploring the viral world through metagenomics. *Curr Opin Virol* 2011;**1**:289–97.

Roux S, Faubladier M, Mahul A, *et al*. Metavir: a web server dedicated to virome analysis. *Bioinformatics* 2011;**27**:3074–5.

Saito R, Smoot ME, Ono K, *et al*. A travel guide to Cytoscape plugins. *Nat Methods* 2012;**9**:1069–76.

Sander M, Schmieger H. Method for host-independent detection of generalized transducing bacteriophages in natural habitats. *Appl Environ Microb* 2001;**67**:1490–3.

Sano E, Carlson S, Wegley L, *et al*. Movement of Viruses between Biomes. *Appl Environ Microb* 2004;**70**:5842–6.

Säwström C, Granéli W, Laybourn-Parry J, *et al*. High viral infection rates in Antarctic and Arctic bacterioplankton. *Environ Microbiol* 2007;**9**:250–5.

Shtarkman YM, Koçer ZA, Edgar R, *et al*. Subglacial Lake Vostok (Antarctica) accretion ice contains a diverse set of sequences

from aquatic, marine and sediment-inhabiting bacteria and eukarya. *PLoS One* 2013;**8**:e67221.

Suttle CA. Marine viruses–major players in the global ecosystem. *Nat Rev Microbiol* 2007;**5**:801–12.

Terns MP, Terns RM. CRISPR-based adaptive immune systems. *Curr Opin Microbiol* 2011;**14**:321–7.

Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* 2000;**45**:1320–8.

Thurber RV, Haynes M, Breitbart M, *et al*. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009;**4**: 470–83.

Vandamme P, Coenye T. Taxonomy of the genus Cupriavidus: a tale of lost and found. *Int J Syst Evol Micr* 2004;**54**: 2285–9.

von Rozycki T, Nies DH. Cupriavidus metallidurans: evolution of a metal-resistant bacterium. *Antonie Van Leeuwenhoek* 2009;**96**:115–39.

Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 1996;**272**:1910–4.

Weinbauer MG. Ecology of prokaryotic viruses. *FEMS Microbiol Rev* 2004;**28**:127–81.

Weinbauer MG, Rassoulzadegan F. Are viruses driving microbial diversification and diversity? *Environ Microbiol* 2003;**6**:1–11.

Wells LE, Deming JW. Brief report Modelled and measured dynamics of viruses in Arctic winter sea-ice brines. *Environ Microbiol* 2006;**8**:1115–21.

Williamson KE, Corzo KA, Drissi CL, *et al*. Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. *Biol Fert Soils* 2013;**49**:857–69.

Xiang S, Yao T, An L, *et al*. 16S rRNA sequences and differences in bacteria isolated from the Muztag Ata glacier at increasing depths. *Appl Environ Microb* 2005;**71**:4619–27.