High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils

Zablocki, O.; van Zyl, Lonnie; Adriaenssens, Evelien M.; Rubagotti, Enrico; Tuffin, Marla; Cary, Stephen Craig; Cowan, Don.

Abstract

The metaviromes of two distinct Antarctic hyperarid desert soil communities have been characterized. Hypolithic communities, cyanobacterium-dominated assemblages situated on the ventral surfaces of quartz pebbles embedded in the desert pavement, showed higher virus diversity than surface soils, which correlated with previous bacterial community studies. Prokaryotic viruses (i.e., phages) represented the largest viral component (particularly Mycobacterium phages) in both habitats, with an identical hierarchical sequence abundance of families of tailed phages (Siphoviridae>Myoviridae>Podoviridae). No archaeal viruses were found. Unexpectedly, cyanophages were poorly represented in both metaviromes and were phylogenetically distant from currently characterized cyanophages. Putative phage genomes were assembled and showed a high level of unaffiliated genes, mostly from hypolithic viruses. Moreover, unusual gene arrangements in which eukaryotic and prokaryotic virus-derived genes were found within identical genome segments were observed. Phycodnaviridae and Mimiviridae viruses were the second-most abundant taxa and more numerous within open soil. Novel virophage-like sequences (within the Sputnik clade) were identified. These findings highlight high-level virus diversity and novel species discovery potential within Antarctic hyperarid soils and may serve as a starting point for future studies targeting specific viral groups.

Introduction

Antarctica is the coldest, driest place on Earth (1). Exposed soil areas comprise approximately 0.4% of the continent’ s surface re mainly located in coastal areas, particularly on the Antarctic peninsula and in the McMurdo Dry Valleys (2). These mineral soils are exposed to a range of “extreme” abiotic factors, including very low temperatures, high soil salinity, low water availability and nutrient levels, high levels of UV radiation, and strong, cold winds descending from glaciers or mountain tops (katabatic). Due to these conditions, the most morphologically distinct soil communities (i.e., type I, II, and III hypoliths) are associated with lithic surfaces (3, 4). Hypolithic

communities, occurring on the ventral surfaces of translucent quartz rocks, have been shown to be mostly composed of phototrophic cyanobacterial species (5). These photoautotroph-dominated communities have been attributed crucial roles within the Antarctic soil ecosystem, such as primary productivity and nitrogen input (6, 7). While the composition of these communities is now reasonably well understood (8-12), the associated viruses, with their potential to influence microbial population dynamics and nutrient cycling via viral lysis (13), have yet to be characterized.

No comprehensive analyses of the collective viral genomic content (i.e., the metavirome) of Antarctic soils have yet been published, with the limited number of reported Antarctic viral metagenomic studies focusing on aquatic systems (14-16) and Antarctic megafauna such as seals (17) and penguins (18). Metaviromic surveys of saline meromictic lakes have shown a high level of diversity of virus-like particles (mostly phages) and several vi- rophages (15, 16). To date, the few studies of viruses in Antarctic soils have all focused on classical phage isolation and lytic induction experiments from culturable bacterial species (19-21). Here we report a comprehensive characterization of virus diversity us- ing a metagenomic approach in Antarctic desert soils, with a focus on the double-stranded DNA (dsDNA) virus composition of two common microhabitats: open surface soils and hypolithic communities.

Materials and methods

Sampling location. Samples were collected from the Miers Valley, Ross Dependency in eastern Antarctica (GPS coordinates, 78° 05.6=S, 163° 48.6=E) during the austral summer period of 2011. For the open- soil sample, 1.5 kg of surface soil (0- to 2-cm depth) was collected from an approximately 1-m2 area at a single location. The hypolith sample consisted of 0.5 kg of hypolith scrapings gathered aseptically from a collection of cyanobacterial-type hypoliths ($n > 50$) from an area of approximately 50 m2. The open-soil sample was recovered from within this area. Sampes were transferred and stored in sterile Whirl-Pak bags (product no. B01445WA; Nasco) at below 0° C in the field and during transport and at -80° C in the laboratory.

Sample processing, DNA extraction, and sequencing. Processing of both types of samples was performed similar to the methods in reference 22. Both the open-soil sample and pooled hypolithic samples were suspended in 3 liters of deionized water and shaken vigorously. The solids were allowed to settle, and the supernatant was decanted. The process was repeated, and both supernatants were mixed. The supernatant was centrifuged at 1,593 X *g* for 10 min (Beckman JA10 rotor), decanted, and passed through a 0.22-μm filter (Stericup [500 ml, 0.22 μm]; catalog no. SCGPU05RE; Millipore). Virus particles were collected from the filtrate by centrifugation in a Beckman JA20 rotor at 43,667 X *g* for6h in autoclaved 30-ml Nalgene polypropylene copolymer (PPCO) tubes (catalog no. 3119-0030). The 6 liters was spun down by

discarding the supernatant from each 30-ml tube (8 tubes in a JA20 rotor) after a round of centrifu- gation and then adding another 30 ml of the extract to the tube. The individual pellets were resuspended in 3 ml successively: the first pellet was resuspended in 3 ml Tris-EDTA (TE) buffer, the liquid was then transferred to the next tube, the pellet was resuspended properly and then transferred to the next tube and so on until all pellets were resuspended. The pellets were treated with DNase I (catalog no. EN0521; Fermentas) and RNase A (catalog no. EN0531; Fermentas) to a final concentration of 0.1 μg/ml at 37° C for 1 h. The presence of bacterial DNA was checked by amplifying the 16S rRNA gene (primers E9F and U1510R [23, 24]) as follows: 1 μl of genomic DNA was mixed with 2.5 μl of each primer (10 mM), 2.5 μl of2 μM deoxynucleoside triphosphates (dNTPs), 2.5 μl of 10X DreamTaq buffer (ThermoFisher Scientific, MA, USA), 1 μl of 10- mg/ml bovine serum albumin (BSA), 0.125 μl DreamTaq polymerase (ThermoFisher Scientific, MA, USA), and Milli-Q water to a total volume of 25 μl. PCR was conducted under the following thermal regime: (i) 5 min at 95° C; (ii) 30 cycles, with 1 cycle consisting of 30 s at 95° C, 30 s at 52° C, and 85 s at 72° C; and (iii) 10 min at 72° C. The virus suspension was treated with proteinase K (Fermentas) at a final concentration of 1 μg/ml at 55° C for 2 h. Seventy microliters of SDS (20%) was added and incubated at 37° C for 1 h. Nucleic acids were purified by performing two rounds of phenol-chloroform-isoamyl alcohol (25:24:1) extraction fol- lowed by chloroform-isoamyl alcohol (24:1) phase separation. DNA was precipitated by the addition of 1/10 volume of sodium acetate (3 M; pH 5.2) and 2 volumes of 100% ethanol and left overnight at 4° C. Samples were centrifuged at 29,000 X *g* for 10 min to pellet the DNA, which was resuspended in 30 μl of TE buffer. The DNA was further cleaned using the Qiagen gel extraction kit (Qiaex II; catalog no. 20021; Qiagen). Ten nano- grams of each sample was then used to perform Phi29 amplification (GenomiPhi HY DNA amplification kit; catalog no. 25-6600-20; GE Healthcare) using the manufacturer' s recommendations. Library preparation included a 10% phiX V3 spike per the manufacturer's instructions (25) with the Illumina Nextera XT library prep kit/MiSeq reagent kit V2. The amplified DNA was sequenced (2X [forward and reverse sequencing] 250-bp reads, ~250-bp average insert size) on the Illumina MiSeq se- quencer platform located at the University of the Western Cape, Cape Town, South Africa.

Sequence data analysis. Sequence reads were curated for quality control and adapter trimmed using CLC Genomics version 6.0.1 (CLC, Den- mark), using the default parameters. Unpaired reads were aligned against each other using Bowtie under default parameters. *De novo* assembly for each read data set was performed with both CLC Genomics and DNASTAR Lasergene SeqMan assembler suite using the default parameters. Reads and contigs were uploaded to the MetaVir (26) version 2 server (http://metavir-meb.univ-bpclermont.fr/) and MG-RAST (27)

(http://metagenomics.anl.gov/) server for virus diversity estimations (data available from these webservers). Taxonomic composition by MetaVir was computed from a BLAST comparison with the RefSeq complete viral genome protein sequence database from NCBI (1 May 2013 release) using BLASTp with a threshold of 10-5 on the E value. Assembled reads were searched for open reading frames (ORFs) and compared to the RefSeq complete viral database (through the MetaVir pipeline) and MG-RAST, which include annotations using the following databases, for functional and organism assignment: GenBank, Integrated Microbial Genomes (IMG), Kyoto Encyclopedia of Genes and Genomes (KEGG), Pathosystems Resource Integration Center (PATRIC), RefSeq, SEED, Swiss-Prot, tremble, and eggnog. The subset of affiliated (i.e., predicted genes with a database match) contigs generated by CLC Genomics were compared to the contigs generated by LaserGene using BLASTx under standard parameters. For the assignment of functional hierarchy, COG (clusters of or- thologous groups), KEGG Orthology (KO), and NOG databases were used. Guanine-plus-cytosine (G+C) content was determined by importing .fasta files into BioEdit (28). The presence of tRNAs in annotated

TABLE 1 Next-generation sequencing metadata, including assembly, annotation, and diversity statistics produced by CLC Genomics and MG-RAST server

Parameter[a]	Value for parameter	
	Open-soil library	Hypolith library
Pre-QC no. of reads	1,622,598	3,771,948
Post-QC no. of reads	1,597,524	3,729,606
Average read length (post-QC)	236.95	241.25
Mean G+C content ± SD (%)	52 ± 12	47 ± 8
No. of reads (%) not assembled into contigs	111,385 (6.97)	274,272 (7.35)
No. of contigs generated	22,237	53,695
Minimum length (nt)	200	200
Maximum length (nt)	177,571	50,044
$N_{25}/N_{50}/N_{75}$ (nt)	865/446/267	945/553/354
% unknown proteins	58.5	81.3
% annotated proteins	39.2	14.5
α-Diversity (total, not limited to viruses)	344.508	1,058.398

[a] QC, quality control; nt, nucleotides.

contigs was assessed with the tRNAscanSE software accessible through http://lowelab.ucsc.edu/tRNAscan-SE/ (29). For prediction of phage life-style and host Gram stain reaction, whole-genome protein sequences of candidate phage genomes

were submitted to the online version of PHACTS (http://www.phantome.org/PHACTS/) (30). Aligned marker genes showing sufficient homology (>150 bp; MetaVir) against the contigs were recovered, and phylogenetic analysis was performed using MEGA5 (http://www.megasoftware.net/). Rooted dendrograms were inferred using the maximum likelihood method with a bootstrap test of 1,000 pseudoreplicates. Phylogenetic analysis for virophage sequences was performed independently from MetaVir. Metavirome virophage amino acid sequences, as well as 9 virophage major capsid protein (MCP) sequences obtained from the NCBI GenBank database were aligned with the online version of MAFFT version 7 (http://mafft.cbrc.jp/alignment/software/). Tree construction was conducted as outlined by Zhou et al. (31).

Accession numbers. These sequence data have been submitted to the DDBJ/EMBL/GenBank databases (sequence read archive [SRA]) under study accession no. SRP038018 (hypolith library) and no. SRP035457 (open-soil library).

Results

Viral diversity estimations. The presence of bacterial contamination was deemed negligible in both metavirome libraries (using the 16S gene fragment), as no discernible bands of amplified products were obtained. MiSeq reads ranged from 236 to 241 bp, with an overall higher G+C content within reads obtained from the open-soil library. Sequencing metadata, assembly metrics, and BLASTp searches are summarized in Table 1. BLASTx comparison of contig data sets from both habitat libraries (generated by two separate assemblers) revealed that 99.41% (hypolith library) and 99.5% (open-soil library) of affiliated contigs were shared between the two assembled read data sets. Contigs from CLC Genomics were used for the remainder of the analysis. Aligned against each other, libraries contained 66.01% of reads that were unique to each habitat, while 33.99% were shared (a read aligned at least once). In both read data sets, bacteria were the most rep- resented hits (80.7 to 94.5%). However, these estimations varied depending on the metagenomic platform used and whether reads or contigs were submitted. BLASTp searches of the MetaVir server using contigs produced significantly more virus-related hits compared to searches of the MG-RAST server. For example, 1.9% of the open-soil contigs were predicted to be viral in origin in MG-RAST, while MetaVir with the same data set predicted 18.8%. The same was true for the hypolith library, where MG-RAST pre- dicted 12.8% for viruses, while MetaVir predicted 19.2%. *Archaea*, *Eukaryota*, and "other" represented the smallest fraction, while viruses (particularly in hypolith) were second in terms of contig affiliations (Fig. 1). Total diversity (ex-diversity) was computed by MG-RAST using normalized values, since unequal distribution of reads between open-soil and hypolith libraries were obtained.

The hypolith library was 3-fold more diverse than the open-soil library (344.5 versus 1,058.4 species). In contrast, the open soil showed higher taxonomic abundance (species evenness, -y-diversity) com- pared to the hypolith (15,663 versus 11,480; Table 2). *Proteobac- teria* and *Firmicutes* were the most abundant in both libraries, with viruses in hypolith the third-most-abundant organisms.

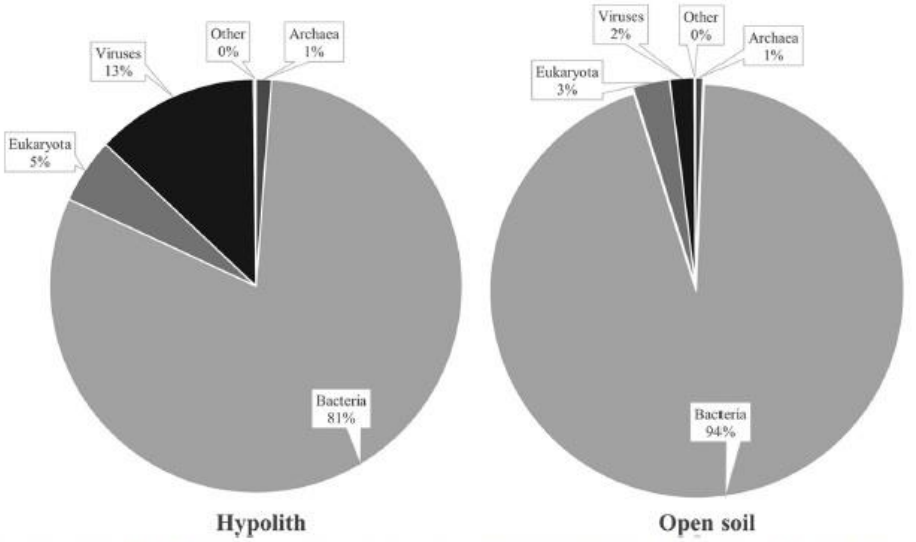

FIG 1 Taxonomic affiliation of the predicted open reading frames (ORFs) in both habitat libraries, assigned by the MG-RAST server.

TABLE 2 Relative abundance of the most represented phyla in both biotopes identified by MG-RAST

Top phylum	Relative abundance[a]		% total abundance[a]	
	OS	HY	OS	HY
Proteobacteria	9,493	3,492	60.6	30.4
Firmicutes	2,722	2,647	17.4	23
Viruses	302	1,469	1.9	12.8
Actinobacteria	521	1,039	3.3	9
Bacteroidetes	1,133	876	7.2	7.6
Cyanobacteria	173	443	1.1	3.9
Chloroflexi	60	121	0.4	1.1
Verrucomicrobia	165	110	1.1	1
Chordata	91	109	0.6	0.9
Unclassified eukaryotes	85	106	0.5	0.9
Planctomycetes	103	87	0.7	0.8

[a] OS, open soil; HY, hypolith.

Rarefaction curves generated by MG-RAST (see Fig. S1 in the supplemental material) showed that the hypolith library was sampled more comprehensively compared to the open-soil library. Nineteen virus families were identified by MetaVir (Table 3), in which prokaryotic viruses were the most abundant in both habitats (76.0% in open soil and 82.3% in hypolith). Identified phages were dominated by the order *Caudovirales* in the following abundance ranking (identical for both habitats): *Siphoviridae* > *Myoviridae* > *Podoviridae*. The next most highly represented virus families were *Mimiviridae* and *Phycodnaviridae*, both more numerous in the open-soil sample. Viral parasites of large dsDNA viruses, i.e., virophages (32), were exclusively identified in the open-soil habitat. Signatures from *Adenoviridae*, *Bicaudaviridae*, *Hytrosaviridae*, *Retroviridae*, and *Rudiviridae* were found in low numbers in the hypolith habitat only. Both habitats contained 13.5 to 15.1% of sequences identified as unclassified viruses.

Due to the lack of universal markers for viruses (such as the 16S rRNA gene marker used for bacteria or the 18S rRNA gene marker for eukaryotes), markers targeting virus families/species were used instead as an alternative to improve taxonomic affiliation of the annotated ORFs from both assembled reads (contigs) and reads alone. Sequences with significant homology to reference markers are shown in Table S1 in the supplemental material. The large terminase subunit (*terL*) marker, required for packaging initiation in members of the *Caudovirales* (33), was the most common match in both habitats. This was consistent with the taxonomic affiliations of contigs in the

virus families shown in Table 3. Non- bacterial viruses (such as *Paramecium bursaria chlorella* virus and *Emiliana huxleyi* virus, which belong to the family *Phycodnaviridae*, and invertebrate viruses belonging to the family *Ascoviridae*), identified with the major capsid protein (MCP) and DNA polymerase family B (*polB*) gene markers, were found exclusively within the open-soil community. For virophage-related sequences, 5 candidate ORFs were submitted to a tBLASTn query

TABLE 3 Taxonomic abundance of identified viral ORFs (BLASTp with a threshold of 10^{-5} for the E value) identified by MetaVir in both Antarctic biotopes

Virus order and family	Host(s)	Taxonomic abundance of viral ORFs (%) (no. of sequence hits)	
		Hypolith	Open soil
Caudovirales			
Myoviridae	Bacteria, archaea	20.9 (1,305)	26.1 (415)
Podoviridae	Bacteria	9.04 (565)	11.0 (175)
Siphoviridae	Bacteria, archaea	52.6 (3,287)	38.3 (610)
Herpesvirales			
Herpesviridae	Vertebrates	0.11 (7)	0.13 (2)
Virus families not assigned into an order[a]			
Adenoviridae	Vertebrates	0.02 (1)	0.00 (0)
Ascoviridae	Invertebrates	0.03 (2)	0.37 (6)
Asfarviridae	Swine, arthropod borne	0.03 (2)	0.06 (1)
Baculoviridae	Invertebrates	0.08 (5)	0.25 (4)
Bicaudaviridae	Archaea	0.05 (3)	0.00 (0)
Hytrosaviridae	Diptera (flies)	0.02 (1)	0.00(0)
Inoviridae	Bacteria	0.16 (10)	0.06 (1)
Iridoviridae	Amphibians, fishes, invertebrates	0.11 (7)	0.44 (7)
Microviridae	Bacteria	0.30 (19)	0.50 (8)
Mimiviridae	Amoebae	0.88 (55)	2.32 (37)
Phycodnaviridae	Algae	1.74 (109)	4.33 (69)
Polydnaviridae	Parasitoid wasps	0.02 (1)	0.00 (0)
Poxviridae	Humans, arthropods, vertebrates	0.06 (4)	0.56 (9)
Retroviridae	Vertebrates	0.02 (1)	0.00 (0)
Rudiviridae	Thermophilic archaea	0.03 (2)	0.00 (0)
Viruses not assigned into families[a]			
Unclassified viruses	N/A[b]	13.5 (844)	15.1 (241)
Sputnik virophage	Mimivirus-infected amoebae	0.00	0.37 (6)

[a] Virus families not yet assigned into an order or family in the International Committee on Taxonomy of Viruses (ICTV) 2012 release (http://ww.ictvonline.org/virusTaxonomy.asp?msl_id=27).
[b] N/A, not available.

and showed closest similarity to the Zamilon (34) and Sputnik (32) virophages, both isolated from soil and aquatic environments, respectively. We attempted to determine its phylogenetic relationship, as among the very few currently recognized virophages, one has been isolated from Organic Lake, Antarctica. Among these ORFs, a partial MCP sequence (344 amino acids long) was identified and aligned with other known virophage MCP sequences (31). The MCP tree in Fig. S2 in the supplemental material shows a clustering pattern identical to the tree in reference 31 and indicates that the virophage sequence from open soil (Miers Valley soil virophage [MVSV]) belonged to the Sputnik virophage group (cluster 1) and was not more closely related to its Antarctic counterpart. Its position within the tree suggests that MVSV shares a genetically distant ancestor with Sputnik and Zamilon virophages.

No reads with significant homology to the *psbA* gene (a marine cyanophage photosynthesis-related gene) were identified. How- ever, other cyanophage sequences were detected within the *g20* and *phoH* phylogenies from the hypolith data set alone (Fig. 2), present as highly divergent sequences at the root of cyanophage sequence clusters. A summary of marker-identified phage species for each marker is shown in Table S2 in the supplemental material. **Functional composition of hypoliths and open soil.** The hypolith data set was highly uncharacterized (predicted proteins with no significant homologs), with 81.3% compared to the open soil with 58.5%. Twenty-six functional categories were assigned to both libraries (Fig. 3), each subdivided into distinct subsystems. Apart from the phage category, functional abundance in all categories was greater in open soil. Highest abundance variations be- tween both biotopes included several metabolic pathways involving phosphorus, nitrogen, aromatic compounds, and iron metabolism. Dormancy and sporulation-related functions were also notably higher in the open soil. A similar trend was found for stress-related functions, including oxidative, osmotic, and acid stress. However, found almost exclusively in the hypolith library were desiccation stress-related protein functions. Virus-specific functional components were retrieved manually from the Meta-Vir server, counted, and classified into several virus component categories, shown in Table S3 in the supplemental material. Both habitat samples contained genes encoding numerous virus structural proteins (portal, tape measure, and capsid) and enzymes (terminases, DNA/RNA polymerases, helicases, and lysins), consistent with an abundance of tailed-phage-related components.

Due to the large number of assembled contigs, a subset of 26 were selected for further analysis (see Table S4 in the supple- mental material) based on a combination of criteria: size (≥10,000 bp), percentage of annotated ORFs within a contig (11 to 100%), and predicted circularity of the putative genome.

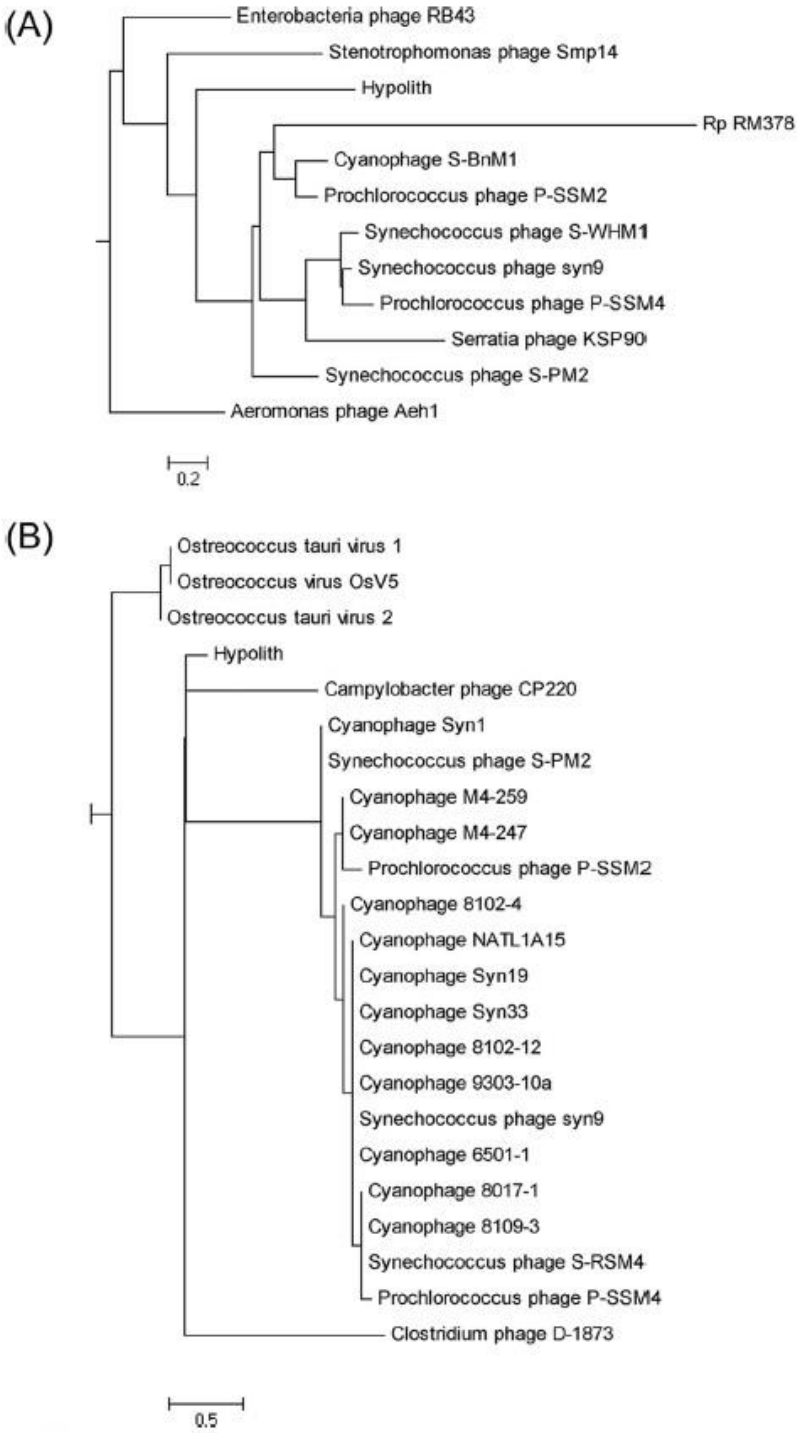

FIG 2 Selected cyanophage subtree phylogenies from *g20* (A) and *phoH* (B) marker genes based on protein alignments retrieved from the MetaVir 2.0 analysis server (metagenomic read selection and tree construction methods outlined by Roux et al. [26]). Scale bars indicate number of base substitutions per site. Rp, *Rhodothermus* phage.

On average, the percentage of homologous genes from public databases was 40.2 ± 21.2 in open soil and 31.9 ± 12.5 in the hypolith. The average values for G+C content in open-soil and hypolith contigs were 55.6% ± 7.3% and 44.3% ± 3.2%, respectively. Phage genomes were submitted to PHACTS (30) for lifestyle (temperate or lytic) and host Gram reaction prediction. As a general trend for both habitats, putative temperate phages dominated (61.5%), while the predicted host range was 88.5% Gram negative. These predictions are supported by a recent study (35), which reported that Gram-negative *Proteo-bacteria* were the dominant phylum in hypolithic and open-soil habitats within the McMurdo Dry Valleys.

For the open-soil habitat alone, contigs contained genes from two virus families infecting algae (*Phycodnaviridae*-like) and amoeba (*Mimiviridae*-like), positioned between phage-related genes. The largest contig (AntarOS_1 [Antar stands for Antarctic, and OS stands for open soil], 177,571 bp) contained one gene from *Acanthamoeba polyphaga* mimivirus and one from *Parame-cium bursaria* chlorella virus A1, while the rest of the genes were phage related. Several core genes (36) from the nucleocytoplasmic large DNA viruses (NCLDVs) were identified in the open soil (also to a lesser extent in the hypolith library) contig data set. These core genes included topoisomerase II, RNA polymerase subunit 2, guanylyltransferase, RuvC, dUTPase 2, thymidylate kinase, MutT/ NUDIX motif, and ankyrin repeat genes. A hybrid gene arrangement from different viruses was found in another contig, AntarOS_17 (Fig. 4). This 24,870-bp contig was divided into 30 predicted ORFs, 21 of which showed signifi homology to virus genes in the RefSeq database (detailed BLAST results for individual ORFs can be found in Table S5 in the supplemental material). Of the 30 predicted ORFs from gene 11 to gene 30 but excluding gene 18, 67% showed significant homology at the amino acid level with a single microalga-infect- ing virus species (unclassifi *Tetraselmis viridis* virus S1, GenBank accession no. NC_020869.1). Genes 4 to 8, 10, and 18 showed similarity to several phages infecting *Burkholderia*, *Rhodobacter*, and *Azospirillum* species. The genome size was too short to be considered phycodnavirus-like (37) and possessed phage genes that were in usual functional synteny toward phage head maturation (such as *terS*, *terL*, and the capsid protein gene [within ORFs 4 to 8]). A tBLASTn search using the protein sequence of the *terL* gene against the NCBI nr database showed highest similarity (34 to 35% identity, E value of 6 X 10-76) to various *Streptococcus* phi phages, including SsUD1, m46.1, and D12. To verify that this contig did not result from read misassembly (i.e., chimeric), two sets of primers were designed to amplify fragments from the over- lapping region between ORF 10 and ORF 11, which based on the contig annotations, appeared to delineate two gene sets from different viruses. Amplicons of expected sizes were obtained and sequenced bidirectionally by Sanger technology. The sequenced fragments aligned with their respective regions, which indicated that this contig region was correctly assembled.

Phage-host associations. As both habitats showed a high level of diversity of phage-related sequences, taxonomic affiliation of the reads (marker gene independent) were categorized according to host and relative sequence abundances in both habitat samples (Fig. 5). Phage sequences identified most closely to host species spanning 5 bacterial phyla: *Firmicutes* (7 bacterial genera), *Proteobacteria* (8 bacterial genera), *Cyanobacteria* (3 bacterial genera), *Bacteroidetes* (*Flavobacterium*), and *Actinobacteria* (4 bacterial genera). By com- paring the bacterial operational taxonomic unit (OTU) distributions in the same soil environments generated by Makhalanyane et al. (35), we attempted to correlate presence/absence of bacterial OTUs based on the phage sequences obtained. Additionally, we included in our comparison 454 sequencing-based soil metagenomic data of Pearce et al. (38), who surveyed moraine soil collected from the margins of a permanent melt water pond located at Mars Oasis on Alexander Island, west of the Antarctic Peninsula. On the basis of identified phage species, only members of *Firmicutes* were found in both soil habitats in this study, but not in the 16S/terminal restriction fragment length polymorphism (TRFLP) bacterial data of Makhalanyane et al. (35). This discrepancy between phage and bacterial data was also observed for hypoliths in hot desert soils (22). The other major bacterial phyla

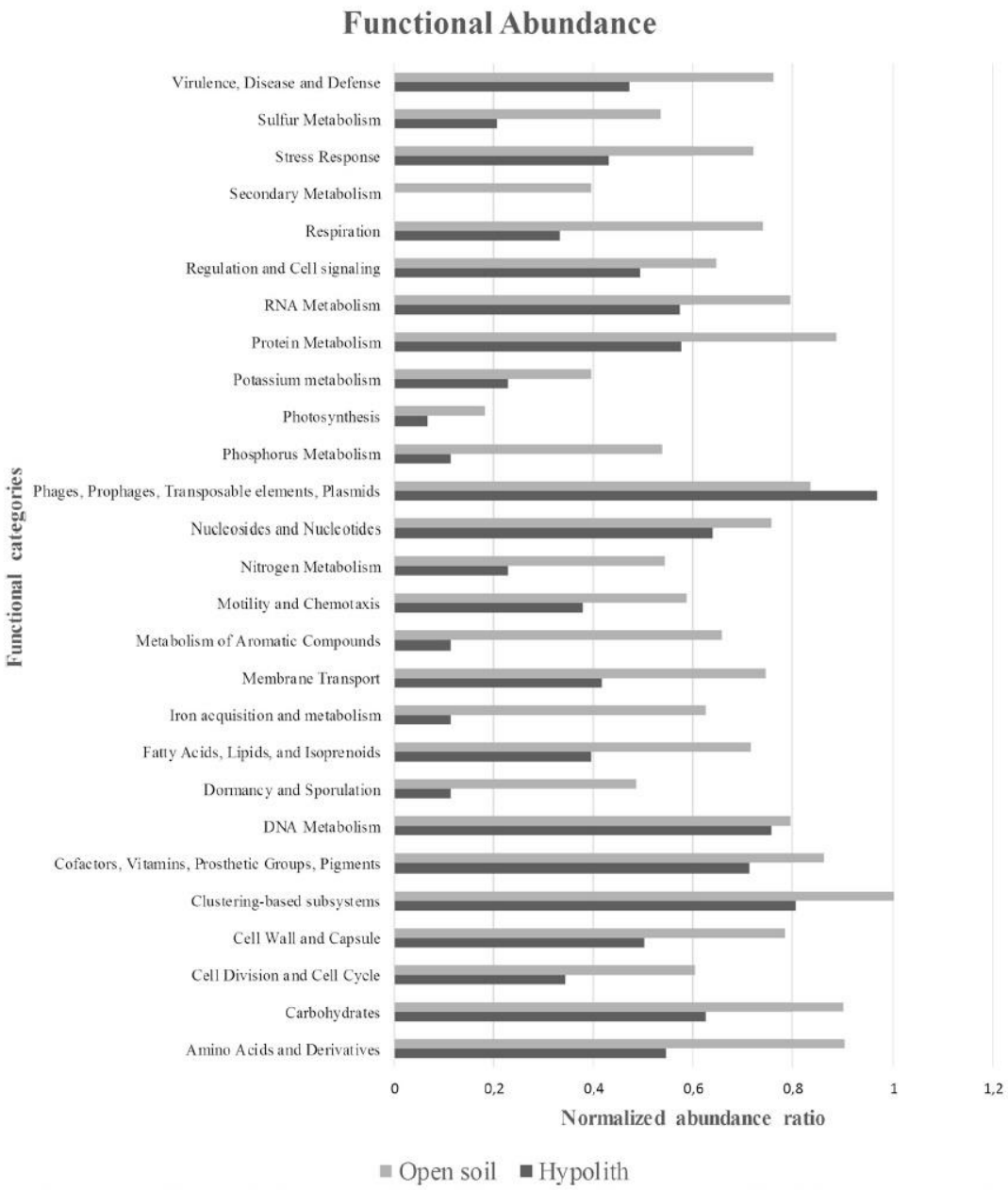

FIG 3 Functional assignment of predicted ORFs compared in both soil habitats. (Functional annotation was performed by MG-RAST using a 60% similarity cutoff.)

(*Proteobacteria*, *Cyanobacteria*, *Bacteroidetes*, and *Actinobacteria*)

were present in both the metavirome data and 16S/TRFLP sequence data. However, bacterial genera indirectly identified by their phages from this study, were all found in the survey by Pearce et al. (38).

At the level of individual phages, *Lactococcus* and *Mycobacte-rium* phage sequences were most common in the hypolith sample (>10%), whereas in open soil, the largest fraction (>6%) was composed of *Bacillus*, *Pseudomonas*, and *Mycobacterium* phage sequences. Few phage host species could be linked to the 16S/ TRFLP data, but at the phylum level, *Proteobacteria* and *Actino-bacteria* were present in both data sets. *Caulobacter* and *Flavobac-terium* were both found in this study (identified by their phage) and 16S/TRFLP data. However, the top 10 virus fraction obtained by Pearce et al. (38) was similar to that found in this study, where *Mycobacterium* phages ranked first (in the case of the hypolith

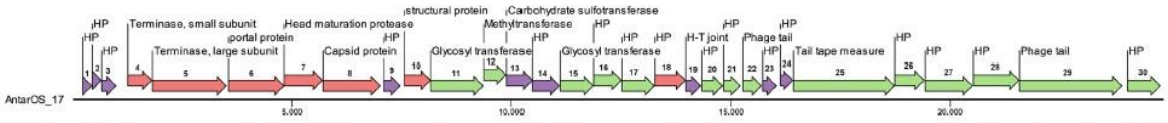

FIG 4 Predicted genome organization of phage AntarOS_17, assembled from open-soil reads. Arrows represent open reading frames (ORFs) and their orientation. ORFs without taxonomic affiliations or ORFs that were predicted without a known function (ORFan) (purple), ORFs that were more closely related to bacteriophage genes (red), and ORFs that were more closely related to *Tetraselmis viridis* virus S1 (a phycodnavirus) (green) are indicated. Numbers below the continuous black line (whole contig length) represent the nucleotide number at a given point. The numbers above each ORF arrow indicate the gene number. HP denotes a hypothetical protein.

sample), but also included *Pseudomonas*, *Enterobacteria*, *Flavo-bacterium*, and *Synechococcus* phages.

Discussion

Unlike aquatic ecosystems which have received considerable attention since the advent of viral metagenomics (39), diversity of virus in many soil habitats has not been characterized extensively. In studies of Antarctic continental microbiology, only freshwater lake metaviromes have been reported thus far (14, 15). Recent phylogenetic studies of Antarctic Dry Valley soils have shown that hypolithic communities represent the most biodiverse and com- plex biological assemblages in this hyperarid soil biome (3). Diversity estimations generated from our data (including viruses) suggest a similar pattern (3-fold-higher diversity compared to open soil). Furthermore, read libraries shared a mere 33% similarity overlap, indicating that hypolithic communities are distinct and differ from their surroundings. This uniqueness, coupled with its higher microbial diversity, may make hypolithic communities biodiversity "micro-hot spots" in this hyperarid desert. A large fraction of ORFs (62.5 to 84.5%) from both soil habitat samples had no significant homologs in public sequence databases, also observed in another published soil metavirome (40). Rarefaction curves indicated that the open-soil biotope has not been sampled sufficiently, and therefore, a greater sequencing depth would be advisable in future metagenome experiments for this habitat.

Taxonomic/functional affiliation of predicted ORFs and gene marker analyses (e.g., using *terL*) were consistent with the conclusion that tailed bacteriophages were the primary virus component in both soil habitats. Furthermore, an identical family-specific hi- erarchical abundance was observed for both habitats (*Siphoviridae*> *Myoviridae* > *Podoviridae*) but with a higher sequence diversity in hypolith communities. Compared to other virus groups, phage sequences were predictably overrepresented in both metaviromes, given that the biotic component in these soils is dominated by prokaryotes (9). However, we note that our nucleic acid extraction method would exclude RNA viruses (either single or double stranded), and therefore, we do not claim that our data reflect the complete viral diversity in these soil habitats. As a sampling bias, viruses in a prophage state may constitute a large (and unsurveyed) proportion of the dsDNA phage diversity, given that it has been reported (19) that many soilborne bacteria appear to contain prophages, including those from Antarctic soils. Conversely, our data suggest that ~61% of phage assemblages are temperate.

Very few archaeal virus signatures were found in either soil habitat, consistent with previous prokaryote diversity studies (9, 35). Unexpectedly, cyanophages were poorly represented in the hypolith sample (in terms of sequence abundance and diversity). Given the dominance of cyanobacteria in type I hypoliths (10, 12), it was reasonably predicted that cyanophages would represent a

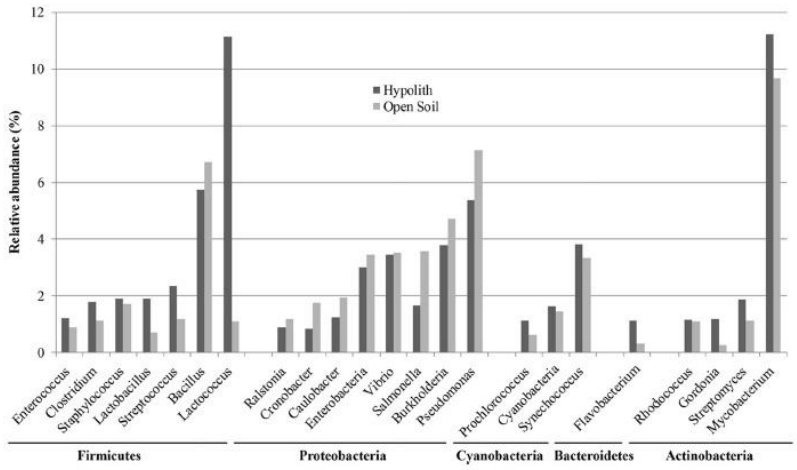

FIG 5 Relative abundance of identified phage isolate sequences based on predicted ORFs identified by MetaVir (E-value cutoff, 10^{-5}) against the RefSeq database. Phage isolates were clustered and counted per their bacterial hosts in their respective habitats (hypolith or open soil). Black lines below a subset of phage-infecting bacterial species indicate the bacterial host phylum.

major clade. The apparent success of cyanobacteria as dominant elements of the hypolithic community might possibly be linked to the low abundance of associated viruses (where the levels of phage infection of other bacterial groups such as mycobacteria, *Bacillus*, *Flavobacterium*, and pseudomonads were higher and their host populations were under tighter predation control). However, this is in contradiction to the general understanding that the most abundant phage groups in any given

environment reflect the abundance of microbial community members found in that environment (41). Thus, while it is possible that cyanophages genuinely represent a minor component of phage diversity, we suggest that this result is an artifact of the substantial underrepresentation of soil-associated cyanophage genome sequences in public data- bases, further accentuated by the fact that most characterized cyanophages are of aquatic origin (42). To our knowledge, only a small fraction of cyanophages of soil origin have been described thus far (43), which also reported a high phylogenetic distance from "common" marine cyanophages, emphasizing the fact that little is known about these cyanophages. In support of this, cyanophage communities in paddy field soils have been shown to be different from those in freshwater, marine water, and even paddy floodwater, identifying unique *g20* subclusters specific to soil-de- rived cyanophages (43). In the current study, marker gene analysis successfully identified several metavirome sequences at the root of cyanophage clusters, suggesting that these represent novel phage phylotypes with a high genetic distance from currently character- ized cyanophages. The high abundance of phages infecting certain bacterial genera such as *Mycobacterium*, *Lactococcus*, *Bacillus*, and *Pseudomonas* phages (~6 to 10% of all identified phage sequences) in both Antarctic desert soil habitats has been reported by Pearce et al. (38).

Sequences with close homology to large dsDNA eukaryotic virus families such as *Mimiviridae* and *Phycodnaviridae*-like genomic elements were found as the second largest virus compo- nent (0.88% to 4.33%) in both habitats (excluding the unclassifiable virus fraction [13.5% to 15.1%]). Mimivirus-related sequences were unexpected, as a 0.22-μm filter size should have excluded large virus particles (~0.7 μm [44]), as well as repeated centrifugation steps. Phycodnaviruses, at ~0.16 ± 0.06 μm (45), would be expected to be recovered in the filtrate. However, detection of MCP components from a novel Sputnik-like virophage, a parasite of large dsDNA viruses (32), provided further indirect evidence for the presence of mimivirus-like populations in the open-soil habitat. In addition, the identified virophage sequence was more closely related to geographically distant isolates (France and Tunisia) compared to the other virophage isolate from Organic Lake in Antarctica. A recent study (38) showed sequences belonging to both host genera (*Paramecium*, *Chlorella*, and *Acan- thamoeba*) and their associated viruses (chlorovirus and mimivirus) in moraine Antarctic soil. La Scola et al. (46) first demonstrated the presence of mimiviruses in soil (previously only isolated from aquatic habitats). The present metavirome se- quences, combined with pyrosequencing data of metagenomic libraries from Pearce et al. (38), provide additional evidence for the presence of mimivirus-like genome elements in Antarctic soils. Further sampling to isolate virophages from Antarctic soils would provide further understanding into the ecology and function of these infectious agents, given that their contributions into the regulation of viral populations are starting to become apparent in other habitats (34). The unusual gene configuration observed

within contig AntarOS_17 (where phage and eukaryotic viruses were predicted) was confirmed by PCR on the original DNA sam- ple, therefore ruling out misassembly of the reads for this region. Most likely, this was caused by misannotation of the predicted ORFs, caused by a lack of closer homologs in databases. While read misassembly is still a possibility in other generated contigs, confidence level in assembly accuracy was high, as the Illumina sequencing control used in both runs was phiX174 (a ~5,000-bp single stranded DNA [ssDNA] virus) which was reassembled almost completely (99.7%) and correctly annotated by the MetaVir pipeline.

Virus families representing less than 0.5% of sequence abundance (Table 3) included those infecting infected Diptera, arthropods, and other invertebrates and were mostly found in the open- soil habitat. As these hosts have been shown to occur on the Antarctic peninsula (47, 48), this may represent an additional pool of uncharacterized viruses within the Antarctic invertebrate fauna.

A positive correlation between phage genera from this study and their associated hosts identified in other bacterial diversity studies was established (35, 38, 49, 50). As in previous hypolith/ open-soil community diversity (ex-diversity) comparisons (Makhalan- yane et al. [35]), where hypoliths showed ahigher degree of diversity than open soil, the same was demonstrated to be true for their associated vi- ruses.

This study represents an initial broad survey of virus diversity in Antarctic hyperarid desert soils and has demonstrated that these local virus assemblages are highly diverse and largely un- characterized. Due to a huge gap in terms of homologous sequences in databases at this time, the generation of additional metagenomic sequence data is not likely to yield usable information. This emphasizes the need for more "traditional" studies, performed in parallel on identical sample sources. These include morphological data from microscopy, lytic induction (e.g., mitomycin C) upon raw soil, and Sanger sequencing of clones targeting specific virus families. Unfortunately, a large fraction will most likely remain uncharacterizable *in vitro*, as the majority of their hosts (bacteria in particular) remain unculturable. Larger eukaryotic viruses infecting algae, amoebae, and invertebrates have not previously been characterized in this environment, and our data demonstrate that these viruses represent an unknown virus population that awaits characterization. Such data would further advance our understanding of the trophic structure and function of communities inhabiting this cold, hyperarid desert biome.

Acknowledgments

We gratefully acknowledge financial support from the National Research Foundation (NRF) (SANAP), the University of Waikato's NZTABS program, Antarctica New Zealand, and the University of Pretoria Genomics Research Institute. We also thank the

Centre for High Performance Computing (CHPC), an initiative supported by the Department of Science and Technology of South Africa. E. M. Adriaenssens holds a Vice-Chancellor' s Fellowship at the University of Pretoria, and O. Zablocki is supported by the NRF-DST Doctoral Innovation Fund.

The opinions expressed and conclusions reached in this article are those of the authors and are not necessarily to be attributed to the NRF.

We declare that we have no conflicts of interest.

References

1. Oerlemans J, Fortuin JPF. 1990. Parameterization of the annual surface temperature and mass balance of Antarctica. Ann. Glaciol. 14:78 – 84.
2. Bockheim JG, McLeod M. 2008. Soil distribution in the McMurdo Dry Valleys, Antarctica. Geoderma 144:43– 49. http://dx.doi.org/10.1016/j.geoderma.2007.10.015.

3. Cary SC, McDonald IR, Barrett JE, Cowan DA. 2010. On the rocks: the microbiology of Antarctic Dry Valley soils. Nat. Rev. Microbiol. 8:129 –138. http://dx.doi.org/10.1038/nrmicro2281.

4. Wierzchos J, de los Ríos A, Ascas C. 2013. Microorganisms in desert rocks: the edge of life on Earth. Int. Microbiol. 15:172–182. http://dx.doi.org/10.2436/20.1501.01.170.

5. Wood SA, Rueckert A, Cowan DA, Cary SC. 2008. Sources of edaphic cyanobacterial diversity in the Dry Valleys of Eastern Antarctica. ISME J. 2:308 –320. http://dx.doi.org/10.1038/ismej.2007.104.

6. Tracy CR, Streten-Joyce C, Dalton R, Nussear KE, Gibb KS, Christian KA. 2010. Microclimate and limits to photosynthesis in a diverse community of hypolithic cyanobacteria in northern Australia. Appl. Environ. Microbiol. 12: 592– 607. http://dx.doi.org/10.1111/j.1462-2920.2009.02098.x.

7. Cowan DA, Sohm JA, Makhalanyane TP, Capone DG, Green TGA, Cary SC, Tuffin IM. 2011. Hypolithic communities: important nitrogen sources in Antarctic desert soils. Environ. Microbiol. Rep. 3:581–586. http://dx.doi.org/10.1111/j.1758-2229.2011.00266.x.

8. Smith MC, Bowman JP, Scott FJ, Line ME. 2000. Sublithic bacteria associated with Antarctic quartz stones. Antarct. Sci. 12:177–184. http://dx.doi.org/10.1017/S0954102000000237.

9. Pointing SB, Chan Y, Lacap DC, Lau MC, Jurgens JA, Farrell RL. 2009. Highly specialized microbial diversity in hyper-arid polar desert. Proc. Natl. Acad. Sci. U. S. A. 106:19964 –19969. http://dx.doi.org/10.1073/pnas.0908274106.

10. Cowan DA, Khan N, Pointing S, Cary SC. 2010. Diverse hypolithic refuge communities in Antarctic Dry Valleys. Antarct. Sci. 22:714 –720. http://dx.doi.org/10.1017/S0954102010000507.

11. Cowan DA, Pointing SB, Stevens MI, Cary SC, Stomeo F, Tuffin IM. 2011. Distribution and abiotic influences on hypolithic microbial com- munities in an

Antarctic Dry Valley. Polar Biol. 34:307–311. http://dx.doi.org/10.1007/s00300-010-0872-2.

12. Khan N, Tuffin M, Stafford W, Cary C, Lacap DC, Pointing SB, Cowan DA. 2011. Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. Polar. Biol. 34:1657–1668. http://dx.doi.org/10.1007/s00300-011-1061-7.

13. Laybourn-Parry J. 2009. No place too cold. Science 324:1521–1522. http://dx.doi.org/10.1126/science.1173645.

14. Säwström C, Lisle J, Anesio AM, Priscu JC, Laybourn-Parry J. 2008. Bacteriophage in polar inland waters. Extremophiles 12:167–175. http://dx.doi.org/10.1007/s00792-007-0134-6.

15. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí

    a. 2009. High diversity of the viral community from an Antarctic lake. Science 326:858 – 861. http://dx.doi.org/10.1126/science.1179287.

16. Yau S, Lauro FM, DeMaere MZ, Brown MV, Raftery MJ, Andrews- Pfannkoch C, Lewis M, Hoffman JM, Gibson JA, Cavicchioli R. 2011. Virophage control of Antarctic algal host-virus dynamics. Proc. Natl. Acad. Sci. U. S. A. 108:6163–6168. http://dx.doi.org/10.1073/pnas.1018221108.

17. Kennedy S, Kuiken T, Jepson PD, Deaville R, Forsyth M, Barrett T, Wilson S. 2000. Mass die-off of Caspian seals caused by canine distemper virus. Emerg. Infect. Dis. 6:637– 639. http://dx.doi.org/10.3201/eid0606.000613.

18. Wallensten A, Munster VJ, Osterhaus ADME, Waldenström J, Bonnedahl J, Broman T, Fouchier RAM, Olsen B. 2006. Mounting evidence for the presence of infl A virus in the avifauna of the Antarctic region. Antarct. Sci. 18:353–356. http://dx.doi.org/10.1017/S095410200600040X.

19. Williamson KE, Radosevich M, Smith DW, Wommack KE. 2007. Inci- dence of lysogeny within temperate and extreme soil environments. Envi- ron. Microbiol. 9:2563–2574. http://dx.doi.org/10.1111/j.1462-2920.2007.01374.x.

20. Meiring TL, Tuffin IM, Cary C, Cowan DA. 2012. Genome sequence of temperate bacteriophage Psymv2 from Antarctic Dry Valley soil isolate Psychrobacter sp. MV2. Extremophiles 16:715–726. http://dx.doi.org/10.1007/s00792-012-0467-7.

21. Swanson MM, Reavy B, Makarova KS, Cock PJ, Hopkins DW, Torrance L, Taliansky M. 2012. Novel bacteriophages containing a genome of another bacteriophage within

their genomes. PLoS One 7:e40683. http://dx.doi.org/10.1371/journal.pone.0040683.

22. Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA. 9 June 2014. Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ. Microbiol. http://dx.doi.org/10.1111/1462-2920.12528.

23. Hansen MC, Tolker-Nielsen T, Givskov M, Molin S. 1998. Biased 16S rDNA PCR amplification caused by interference from DNA flanking the template region. FEMS Microbiol. Ecol. 26:141–149. http://dx.doi.org/10.1111/j.1574-6941.1998.tb00500.x.

24. Reysenbach AL, Pace NR, Robb FT, Place AR. 1995. Archaea: a labora- tory manual—thermophiles. Cold Spring Harb. Protoc. 16:101–107.

25. Illumina, Inc. 2012. Nextera XT DNA sample preparation guide. Part 15031942. Revision B. Illumina, Inc, San Diego, CA.

26. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. 2011. Metavir: a web server dedicated to virome analysis. Bioinformatics 27:3074 –3075. http://dx.doi.org/10.1093/bioinformatics/btr519.

27. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodrigez A, Stevens R, Wilke A, Wilkening J, Edwards RA. 2008. The metagenomics RAST server—a public resource for the auto- matic phylogenetic and functional analysis of metagenomes. BMC Bioin- formatics 9:386. http://dx.doi.org/10.1186/1471-2105-9-386.

28. Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.

29. Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686 – W689. http://dx.doi.org/10.1093/nar/gki366.

30. McNair K, Bailey BA, Edwards RA. 2012. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 28:614 – 618. http://dx.doi.org/10.1093/bioinformatics/bts014.

31. Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, Pan Y, Wang Y. 2013. Diversity of virophages in metagenomic data sets. J. Virol. 87:4225– 4236. http://dx.doi.org/10.1128/JVI.03398-12.

32. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, Raoult D. 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455:100 –104. http://dx.doi.org/10.1038/nature07218.

33. Black LW. 1995. DNA packaging and cutting by phage terminases: con- trol in phage T4 by a synaptic mechanism. Bioessays 17:1025–1030. http://dx.doi.org/10.1002/bies.950171206.

34. Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D, La Scola B. 2014. Zamilon, a novel virophage with Mimiviridae host specificity. PLoS One 9:e94923. http://dx.doi.org/10.1371/journal.pone.0094923.

35. Makhalanyane TP, Valverde A, Birkeland NK, Cary SC, Tuffin IM, Cowan DA. 2013. Evidence for successional development in Antarctic hypolithic bacterial communities. ISME J. 7:2080 –2090. http://dx.doi.org/10.1038/ismej.2013.94.

36. Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J. Virol. 75:11720 –11734. http://dx.doi.org/10.1128/JVI.75.23.11720-11734.2001.

37. Van Etten JL, Graves MV, Müller DG, Boland W, Delaroque N. 2002. Phycodnaviridae—large DNA algal viruses. Arch. Virol. 147:1479 –1516. http://dx.doi.org/10.1007/s00705-002-0822-6.

38. Pearce DA, Newsham KK, Thorne MA, Calvo-Bado L, Krsek M, Laskaris P, Hodson A, Wellington EM. 2012. Metagenomic analysis of a southern maritime Antarctic soil. Front. Microbiol. 3:403. http://dx.doi.org/10.3389/fmicb.2012.00403.

39. Srinivasiah S, Bhavsar J, Thapar K, Liles M, Schoenfeld T, Wommack KE. 2008. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Res. Microbiol. 159:349 –357. http://dx.doi.org/10.1016/j.resmic.2008.04.010.

40. Fierer N, Breitbart M, Nulton J, Salamon P, Loozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB. 2007. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, fungi, and viruses in soil. Appl. Environ. Microbiol. 73:7059 –7066. http://dx.doi.org/10.1128/AEM.00358-07.

41. Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 13:278 –284. http://dx.doi.org/10.1016/j.tim.2005.04.003.

42. Hurst CJ (ed). 2011. Studies in viral ecology: microbial and botanical host systems, vol 1. Wiley-Blackwell, Hoboken, NJ.

43. Wang G, Asakawa S, Kimura M. 2011. Spatial and temporal changes of cyanophage communities in paddy field soils as revealed by the capsid assembly protein gene g20. FEMS Microbiol. Ecol. 76:352–359. http://dx.doi.org/10.1111/j.1574-6941.2011.01052.x.

44. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J,

45. Abergel C. 2009. Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. J. Invertebr. Pathol. 101:172–180. http://dx.doi.org/10.1016/j.jip.2009.03.011.
46. Dunigan DD, Fitzgerald LA, Van Etten JL. 2006. Phycodnaviruses: a peek at genetic diversity. Virus Res. 117:119 –132. http://dx.doi.org/10.1016/j.virusres.2006.01.024.

47. La Scola B, Campocasso A, N=Dong R, Fournous G, Barrassi L, Flau- drops C, Raoult D. 2010. Tentative characterization of new environmen- tal giant viruses by MALDI-TOF mass spectrometry. Intervirology 53: 344 –353. http://dx.doi.org/10.1159/000312919.

48. Convey P, Gibson JA, Hillenbrand CD, Hodgson DA, Pugh PJ, Smellie

49. JL, Stevens MI. 2008. Antarctic terrestrial life— challenging the history of the frozen continent? Biol. Rev. Camb. Philos. Soc. 83:103–117. http://dx.doi.org/10.1111/j.1469-185X.2008.00034.x.
50. Treonis AM, Wall DH, Virginia RA. 1999. Invertebrate biodiversity in Antarctic Dry Valley soils and sediments. Ecosystems 2:482– 492. http://dx.doi.org/10.1007/s100219900096.

51. Friedmann EI (ed). 1993. Antarctic microbiology, p 634. Wiley-Liss, New York, NY.

52. Smith JJ, Tow LA, Stafford W, Cary C, Cowan DA. 2006. Bacterial diversity in three different Antarctic cold desert mineral soils. Microb. Ecol. 51:413– 421. http://dx.doi.org/10.1007/s00248-006-9022-3.