Niche-dependent genetic diversity in Antarctic metaviromes

Olivier Zablocki[1,2], Lonnie van Zyl[3], Evelien M. Adriaenssens[1,2], Enrico Rubagotti[2], Marla Tuffin[3], Stephen C. Cary[4], Don Cowan[1,2]

[1]Centre for Microbial Ecology and Genomics, University of Pretoria, Pretoria, South Africa;

[2]Genome Research Institute, University of Pretoria, Pretoria, South Africa;

[3]Institute for Microbial Biotechnology and Metagenomics, University of the Western Cape, Bellville, South Africa;

[4]The International Centre for Terrestrial Antarctic Research, University of Waikato, Hamilton, New Zealand

Keywords: Antarctica, metaviromics, *Caudovirales*, temperate phage, virophage, hypolith.

Citation of original article:

Zablocki, O., van Zyl, L., Adriaenssens, E. M., Rubagotti, E., Tuffin, M., Cary, C., & Cowan, D. (2014). High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. *Appl. Environ. Microbiol*., doi:10.1128/AEM.01525-14.

Abstract

The metaviromes from two different Antarctic terrestrial soil niches have been analysed. Both hypoliths (microbial assemblages beneath transluscent rocks) and surrounding open soils showed a high level diversity of tailed phages, viruses of algae and amoeba, and virophage sequences. Comparisons of other global metaviromes with the Antarctic libraries showed a niche-dependent clustering pattern, unrelated to the geographical origin of a given metavirome. Within the Antarctic open soil metavirome, a putative circularly permuted, ~42kb dsDNA virus genome was annotated, showing features of a temperate phage possessing a variety of conserved protein domains with no significant taxonomic affiliations in current databases.

Introduction

The hyperarid soils of the Antarctic Dry valleys were for long thought to harbour low numbers of microorganisms.[1] However, molecular tools such as 16S rRNA analysis have provided a more realistic view of the true microbial diversity within this polar desert ecosystem.[2,3] Cyanobacterial communities, in particular, have been attributed key roles within this ecosystem[4], and are commonly found on the ventral surface of translucent rocks, termed hypoliths.[5,6] While these have received much attention and have now been well characterized in terms of taxonomic diversity[7], they do not necessarily form the basal tier of the food chain, as associated bacteriophages may be involved in the regulation, survival and evolution of these communities. In our recent paper, the viral component of cyanobacterial-dominated bacterial communities associated with quartz rocks (i.e. Type I hypoliths) and the surrounding open soils were characterized using a shotgun metagenomic approach. A high diversity of viruses was found in both habitats, while cyanophage marker genes were poorly

represented. In this addendum, we present a global comparison of the hypolith and open soil metaviromes, using additional publicly available data from a range of habitats, including a hot hyperarid biome, the Namib Desert. We also present a complete, circular, dsDNA temperate phage genome.

Globally related, niche-specific microbial communities

The hypolith metavirome showed greater viral diversity than the open soil metavirome, whereas the latter contained greater sequence abundance and lower sequence diversity. However, rarefaction curves suggested that sequence diversity may be under-estimated. Reads from both habitat libraries shared a 33% sequence identity overlap, indicating very genetically distinct communities despite their close habitat proximity.

A BLASTp-based comparison (10^{-5} threshold on the E-value) between hypolith and open soil metaviromes has already been discussed in detail in our recent paper. However, our dataset contained a large number of unknown/unaffiliated sequences (58.5% and 81.3 % for open soil and hypolith samples, respectively). For such datasets, whole metavirome nucleotide frequency (di-, tri- and tetranucleotide) comparisons can be a valuable tool for assigning putative ecological classifications, without the requirement of homology against reference databases.[8] Contig datasets from several metaviromes from a range of habitats (freshwater[9], seawater[10], plant-associated[11], Namib Desert open soil (unpublished data) and hypoliths[12], available from the MetaVir server[13]) were selected for dinucleotide frequency comparisons[8] with the Antarctic sequence datasets.

Figure 1 shows that the 9 metaviromes clustered in two separate groups. Group 1, composed of plant and water metaviromes, were further sub-divided into freshwater and seawater clades. Group 2 was composed of soil-associated habitats, sub-divided

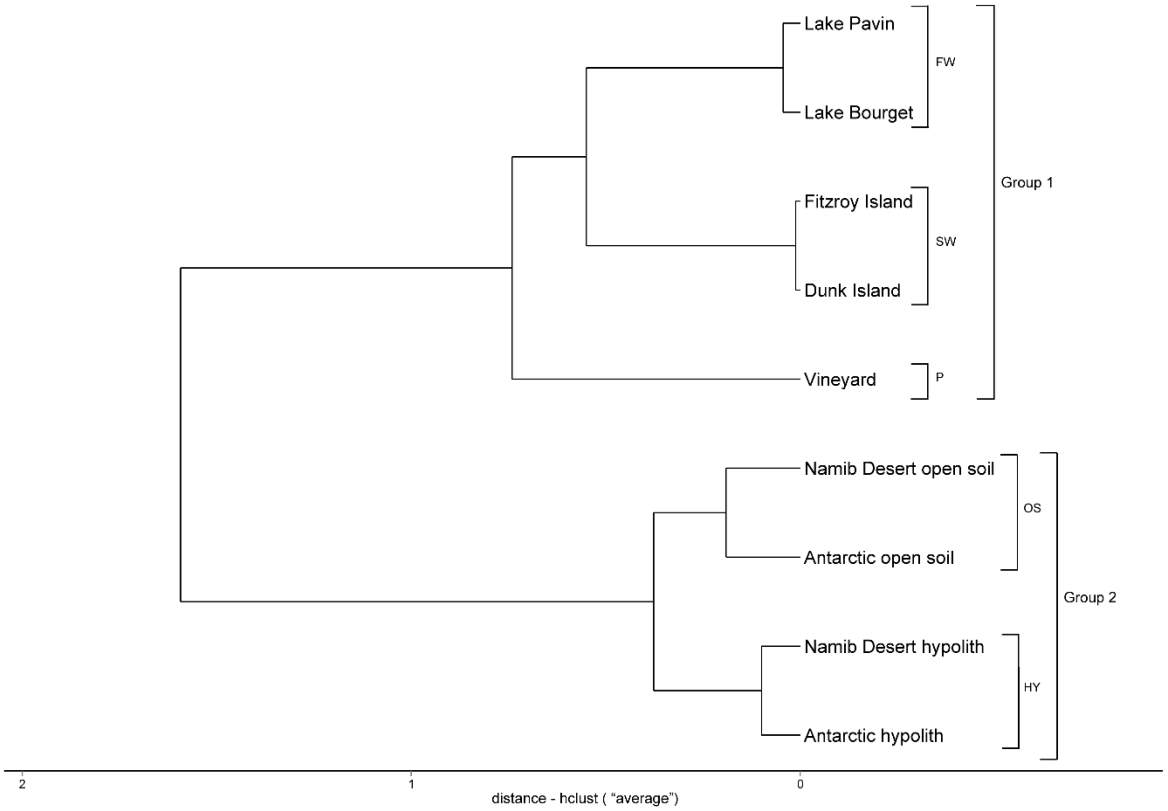

Figure 1: Hierarchical clustering of various metaviromes (assembled into contigs) based on dinucleotide frequencies. The x-axis denotes eigenvalues distances. The tree was constructed using MetaVir server pipeline according to the method of Willner et al.8 FW: freshwater; SW: seawater; P: plant; OS: open soil; HY: hypolith

into open soil and soil-associated rock (i.e. hypolith) clades. Interestingly, despite their widely differing habitat-associated environmental characteristics and equally great spatial separation (i.e., Antarctic and Namib Desert), both hypolithic viromes clustered at a single node. The same was true for the open soil metaviromes, with the Antarctic dataset clustered with the Namib Desert dataset. The implications of this result are that the widely differing environmental temperatures do not define the genetic relatedness of the communities, but that aridity may be a strong driver of host and viral diversity.

In both habitats, the majority of predicted virus genes with assigned taxonomy belonged to the order *Caudovirales*. This was expected, as both habitats have been previously shown to be dominated by typical soil prokaryotic taxa. However, several virus families displayed variation in sequence abundance according to niche. In the open soil sample, invertebrate viruses (*Ascoviridae*, *Baculoviridae*, *Iridoviridae*) and large dsDNA viruses (*Mimiviridae*, *Phycodnaviridae*) showed higher abundance than in hypoliths. These results suggest that these viral hosts are less established in hypolithic habitats. Sequences closely related to virophages were found exclusively in the open soil sample. To date, all but one virophage have been isolated from aquatic habitats[14], with this study representing only the second report of such a sub-viral entity in soil. Despite the deep sequencing approach employed in the study, the low number of virophage-related sequences identified would suggest that virophages were either present in low abundance or were poorly enriched by the protocol employed.

Metagenomic assembly of a circularly permuted, temperate phage genome

In our recent paper, a deeper analysis of a subset of annotated contigs was presented, which showed that genes associated with phylogenetically distant virus families occurred within continuous nucleotide sequences (e.g. AntarOS_17). Here we

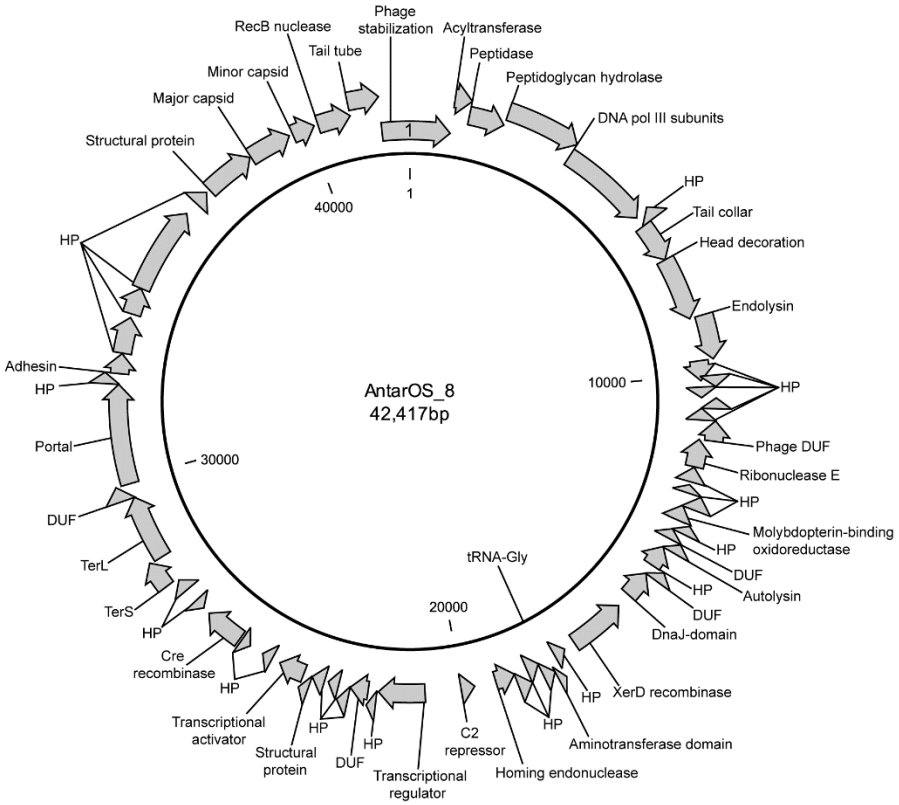

Figure 2: Gene annotation of contig AntarOS_8. Arrowed blocks are open reading frames (ORFs) and their orientation. HP denotes hypothetical proteins and DUF denotes conserved protein domains of unknown function. Numbers within the circle are nucleotide positions, starting within gene number 1 (indicated within the top arrow as "1") and onwards in a clockwise orientation.

describe another annotated contig (assembled with CLC Genomics/DNASTAR Lasergene NGen, method outlined in Zablocki et al.[15]), from the open soil read dataset, in the form of a circularly permuted sequence.

Gene prediction with MetaGeneAnnotator on the MetaVir pipeline yielded 62 open reading frames (ORFs) along the 42,417 base-pair (bp) contig (named AntarOS_8, Figure 2). Fifty-seven percent of the ORFs identified had a database homolog using BLASTp against the non-redundant (nr) GenBank database or HHPred (http://toolkit.tuebingen.mpg.de/hhpred) with HMM-HMM scans against several protein databases (PDB, SCOP, pfamA, smart, PANTHER, TIGRfam, PIRSF and CD) . Gene functions and/or taxonomic affiliations for these putative ORFs are summarized in Table 1.

The mean G+C content of the putative genome was 65.4%. Both ARAGORN[16] and tRNAscan-SE[17] were used to search for tRNAs and tmRNAs. One tRNA gene was identified (tRNA-Gly, anticodon: TCC, 73bp). The gene was positioned at nucleotide position 18,059-18,131, corresponding to a region upstream of gene 27 (position 15,833-17,173), an integrase domain, which together may be involved in the integration of the phage genome by site-specific host genome integration. A XerD tyrosine recombinase domain within gene 45 (similar to phage P1 Cre recombinase[18]) and a putative transcription regulator (gene 34) with a helix-turn-helix (HTH) xenobiotic response element domain (containing 3 non-specific and 6 specific DNA-binding sites), was most similar to the generic structure of the prophage repressor family of transcriptional regulation proteins.[19] This composition of genes strongly suggests that this putative virus genome is that of a temperate bacteriophage.

The presence of the small and large terminase subunit protein domains (gene 48 and 49, respectively), a tail tape measure domain (gene 37), a portal protein (gene

Table 1. Detailed ORF list including peptide length, associated taxonomy and predicted function. "TM" denotes the detection of a transmembrane domain within the predicted protein; a.a: amino acid

ORF	Size (a.a)	Best BLASTp taxonomic hit	E- value	Accession	% identity; % query coverage	Conserved protein domain(accession)	Predicted function
1	566	*Sulfitobacter* phage NYA-2014a	5e-118	AIM40653.1	40; 99	Phage_stabilize (pfam11134)	Phage stabilization protein
2	127	*Burkholderia* phage Bcep22	3e-06	NP_944298.1	31; 100	Ribosomal-protein-alanine acetyltransferase (2z10_A)	Acyl-CoA N-acyltransferase
3	274	*Sulfitobacter* phage PhiCB2047-C	5e-54	YP_007675267.1	44; 91	—	Peptidase domain-containing hypothetical protein
4	557	*Brucella* phage Tb	2e-04	YP_007002033.1	33; 23	—	Peptidoglycan hydrolase
5	705	—	—	—	—	DNA polymerase III subunits gamma and tau (PRK14954)	Gene expression
6	96	—	—	—	—	—	Hypothetical protein
7	307	*Brucella* phage S708	1e-14	AHB81257.1	29; 70	Phage Tail Collar Domain (pfam07484)	Collar protein for virion assembly
8	493	Prophage MuMc02	0.005	ADD94511.1	34; 25	—	Head decoration protein
9	332	Stenotrophomonas phage Smp131	40;39	YP_009008370.1		Lysozyme_like domain (cl00222)	Endolysin (TM)
10	153	—	—	—	—	Phage_HK97_TLTM (PF06120)	Tail length tape measure protein
11	84	—	—	—	—		Hypothetical protein
12	81	—	—	—	—	Orn_Arg_deC_N: Pyridoxal-dependent decarboxylase (PF02784)	Hypothetical protein (TM)
13	79	—	—	—	—	—	Hypothetical protein
14	90	—	—	—	—	—	Hypothetical protein
15	149	*Vibrio* phage pYD38-A	2e-19	YP_008126176.1	42; 65	Phage gp49 66 superfamily (cl10351)	Unknown function
16	214	Mycobacteriophage Macn Cheese	0.010	AFN37792.1	38; 35	ribonuclease E (PRK10811)	Site-specific RNA endonulcease
17	115					—	Hypothetical protein
18	64	—	—	—	—	—	Hypothetical protein
19	105	—	—	—	—	—	Hypothetical protein
20	120	—	—	—	—	MopB_4 CD (cd02765)	Molybdopterin-binding oxidoreductase-like domains
21	71	—	—	—	—		Hypothetical protein
22	77	—	—	—	—	cas_TM1812 CRISPR-associated protein (TIGR02221)	Unknown function
23	76	—	—	—	—	Autolysin_YrvJ (PIRSF037846)	N-acetylmuramoyl-L-alanine amidase
24	166	—	—	—	—	—	Hypothetical protein
25	80					AAA (cl18944)	Unknown function
26	213	Archaeal BJ1 virus	8e-08	YP_919032.1	28; 92	DnaJ domain (cd06257)	DnaJ-containing protein
27	446	—	—	—	—	XerD Site-specific recombinase (COG4974)	Integrase/ recombinase

ORF	Size (a.a)	Best BLASTp taxonomic hit	E- value	Accession	% identity; % query coverage	Conserved protein domain(accession)	Predicted function
28	90	—	—	—	—	—	Hypothetical protein
29	57	—	—	—	—	4-aminobutyrate aminotransferase (PRK06058)	Unknown function
30	88	—	—	—	—	—	Hypothetical protein
31	98	—	—	—	—	—	Hypothetical protein
32	103	*Pseudomonas* phage DMS3	7e-15	YP_950436.1	40; 87	—	Hypothetical protein
33	153	—	—	—	—	HNH endonuclease (PF01844)	HNH homing endonuclease
34	106	—	—	—	—	P22 C2 repressor (d2r1jl_)	C2 repressor
35	360	*Lactobacillus* phage c5	1e-05	YP_007002377.1	32; 27	DnaA N-terminal domain (cl13142); Arsenical Resistance Operon Repressor (smart00418)	Transcriptional regulator
36	67	—	—	—	—	—	Hypothetical protein
37	140	*Azospirillum* phage Cd	9e-24	YP_001686851.1	62; 52	Domain of unknown function (pfam10073)	Conserved hypothetical protein
38	80	—	—	—	—	—	Hypothetical protein
39	83	—	—	—	—	—	Hypothetical protein
40	100	—	—	—	—	—	Hypothetical protein
41	70	—	—	—	—	RHH_1: Ribbon-helix-helix protein, copG family (PF01402)	Structural protein
42	203	—	—	—	—	rfaH transcriptional activator (PRK09014)	Transcriptional activator protein
43	82	—	—	—	—	—	Hypothetical protein
44	79	—	—	—	—	—	Hypothetical protein
45	308	Enterobacteria phage D6	1e-12	AAV84949.1	25; 88	Cre recombinase, C-terminal catalytic domain (cd00799)	Cre recombinase
46	82	—	—	—	—	—	Hypothetical protein
47	98	—	—	—	—	—	Hypothetical protein
48	211	*Sinorhizobium* phage phiLM21	2e-12	AII27789.1	31; 82	Terminase_2 Terminase small subunit (pfam03592)	TerS
49	506	*Xanthomonas* phage Xp15	1e-94	YP_239275.1	41; 90	Terminase_3 Phage terminase large subunit (pfam04466)	TerL (TM)
50	110	*Sulfitobacter* phage PhiCB2047-C	2e-23	YP_007675320.1	49; 85	Protein of unknown function (DUF3307)	Conserved hypothetical protein
51	774	*Brucella* phage Tb	2e-163	YP_007002020.1	40; 88	Portal protein (3lj5_A)	Portal protein
52	82					—	Hypothetical protein
53	155	*Rhizobium* phage vB_RleM_P10VF	3e-09	AIK68325.1	34; 60	Tfp pilus assembly protein (COG3419)	tip-associated adhesin PilY1
54	281	—	—	—	—	—	Hypothetical protein
55	208	—	—	—	—	—	Hypothetical protein
56	678	—	—	—	—	—	Hypothetical protein

(*continued on next page*)

ORF	Size (a.a)	Best BLASTp taxonomic hit	E- value	Accession	% identity; % query coverage	Conserved protein domain(accession)	Predicted function
57	115	—	—	—	—	—	Hypothetical protein
58	400	*Brucella* phage Tb YP_007002024.1	3e-12 25; 90	—	Structural protein		
59	339	*Brucella* phage Tb	7e-60	YP_007002025.1	37; 94	Major capsid protein gp5 (d2ft1a1)	Major capsid protein
60	178	*Myxococcus* phage Mx8 NP_203463.1	9e-24 46; 80		Major virion structural protein		
61	258	*Brucella* phage Pr	1e-17	YP_007002027.1	32; 92	RecB family nuclease (TIGR03491)	DNA-binding protein
62	233	*Brucella* phage Tb	5e-33	YP_007002028.1	35; 93	Tail tubular protein A (PHA00428)	Tail tube protein

51) and a tail collar protein domain (gene 7) together support a tailed phage virion structure. A BLASTp search using the *terL* gene showed closest homology (41% identity with 90% sequence coverage, E-value 1e-94) to *Xanthomonas* phage Xp15 (accession YP_239275.1). This phage isolate belongs to the order *Caudovirales*, but has not been classified at family or genus level. BLASTp analysis of Xanthomonas phage XP15 AAX84861.1 (putative tail tape measure protein) revealed its closest homologs belongs to the family *Siphoviridae*, including *Burkholderia* phage BcepGomr. Similarly, the *terL* gene phylogeny generated by a BLASTp search indicated that the closest homologs belonged to siphoviruses (data not shown).

Several genes on contig AntarOS_8 could not be assigned to any known viral genome, but nevertheless contained conserved sequences translating to proteins with homologs in current databases. These included a site-specific RNA endonuclease (RecA-like, gene 16) and a J-domain (DnaJ) at the C-terminus of a larger peptide of unknown function (possible role in protein folding and degradation).

Gene 35 may be involved in host regulation of expression of non-essential genes. The predicted peptide contained two domains, an arsenic- resistance operon repressor (HTH_ARSR) and a DnaA-like region, probably used for DNA binding and regulation of expression.

Several putative host defence evasion/lysis-related genes were also identified in the AntarOS_8 contig. Gene 32 encoded a putative host nuclease inhibitor and gene 34 showed high homology to a *Sulfitobacter* phage gene encoding an exodeoxyribonuclease. Two genes were predicted to be host cell lysis-related protein (putative peptidoglycan hydrolase (gene 4) and a lysozyme domain superfamily (gene 9)).Gene 53 contained a conserved domain for the Tfp pilus assembly protein, associated with the adhesion tip PilY1. This gene may be used for lysogenic

conversion.

The remainder of predicted proteins were genes coding for an unknown structural protein, an internal virion protein, the major capsid protein(MCP), acyl-CoA-N-acyltransferase, AAT superfamily (PLP-dependent) and hypothetical protein related to diverse phage isolates infecting *Brucella*, *Burkholderia*, *Sulfitobacter*, *Pseudomonas*, *Bacillus*, *Mycobacterium*, *Myxococcus* and *Rhizobium*.

Several of the conserved genes of contig AntarOS_8 show a distant relation to the siphoviruses lambda (integrase (gene 27) and HK97 (MCP) and the podovirus P22 (*terL,* c2). This indicates that this contig represents a novel mosaic phage genome which is a potential new member of the hybrid supercluster of Lambda-like phages recently described by Grose and Casjens and it is speculated that the host belongs to the family *Enterobacteriaceae*[20].

Conclusions

Metaviromic analyses of Antarctic hyperarid niche habitats has revealed that these are highly novel, but not necessarily geographically distinct. Dinucleotide clustering of the metavirome contigs grouped hypolith niche habitats together, as well as open soil, rather than the cold versus warm hyperarid environments. Detailed analysis of a specific contig revealed a distant lambda-like genome, linking it in with a highly diverse group of temperate phages isolated from all over the world.

Acknowledgments

We wish to thank the National Research Foundation (South Africa) and the Genomics Research Institute of the University of Pretoria (South Africa) for financial support. The opinions expressed and conclusions reached in this article are those of the authors and are not necessarily to be attributed to the NRF.

References

1. Wynn-Williams DD. Antarctic microbial diversity: the basis of polar ecosystem processes. Biodivers. Cons. 1996; 5: 1271-1293.
2. Pointing SB, Chan Y, Lacap DC, Lau MC, Jurgens JA, Farrell RL. Highly specialized microbial diversity in hyper-arid polar desert. Proc. Natl. Acad. Sci. U.S.A. 2009; 106: 19964-19969.
3. Makhalanyane TP, Valverde A, Birkeland NK, Cary SC, Tuffin IM, Cowan DA. Evidence for successional development in Antarctic hypolithic bacterial communities. ISME J. 2013; 7: 2080-2090.
4. Cowan DA, Sohm JA, Makhalanyane TP, Capone DG, Green TGA, Cary SC, Tuffin IM. Hypolithic communities: important nitrogen sources in Antarctic desert soils. Environ. Microbiol. Rep. 2011; 3: 581-586.
5. Wood SA, Rueckert A, Cowan DA, Cary SC. Sources of edaphic cyanobacterial diversity in the Dry Valleys of Eastern Antarctica. ISME J. 2008; 2: 308-320.
6. Smith MC, Bowman JP, Scott FJ, Line MA. Sublithic bacteria associated with Antarctic quartz stones. Antarct. Sci. 2000; 12: 177-184.
7. Khan N, Tuffin M, Stafford W, Cary C, Lacap DC, Pointing SB, Cowan D. Hypolithic microbial communities of quartz rocks from Miers Valley, McMurdo Dry Valleys, Antarctica. Polar Biol. 2011; 34: 1657-1668.
8. Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes. Environ. Microbiol 2009; 11: 1752-1766.
9. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Debroas D. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PloS one 2012; 7: e33641.

10. Hurwitz BL, Sullivan MB. The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS One 2013, 8: e57355.
11. Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJG, Burger JT. Deep sequencing analysis of viruses infecting grapevines: virome of a vineyard. Virol. 2010; 400: 157-163.
12. Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan, DA. Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ. Microbiol. 2014.
13. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. Metavir: a web server dedicated to virome analysis. Bioinformatics 2011; 27: 3074-3075.
14. Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, Colson P, Raoult D, La Scola, B. Zamilon, a Novel Virophage with Mimiviridae Host Specificity. PloS one 2014; 9: e94923.
15. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary C, Cowan D. High-level diversity of tailed Phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of Antarctic soils. Appl. Environ. Microbiol. 2014; doi:10.1128/AEM.01525-14.
16. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. 2004; 32: 11-16.
17. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25: 0955-964.

18. Guo F, Gopaul DN, Van Duyne, GD. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature 1997; 389: 40-46.
19. Wood HE, Devine KM, McConnell DJ. Characterisation of a repressor gene (xre) and a temperature-sensitive allele from the Bacillus subtilis prophage, PBSX. Gene 1990; 96:83-88.
20. Grose J, Casjens S. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family *Enterobacteriaceae*. Virology 2014; 468-470:421-443.