www.nature.com/scientificreports

scientific reports

OPEN

Novel adaptive immune systems in pristine Antarctic soils

Marc W. Van Goethem[1,2], Oliver K. I. Bezuidt[3,4], Rian Pierneef[3,4], Surendra Vikram[1], David W. Hopkins[5], Thomas Aspray[6], Grant Hall[7], Stephan Woodborne[8], Ian D. Hogg[9], Trent R. Northen[10], Weidong Kong[11], Daniele Daffonchio[2], Don A. Cowan[1], Yves Van de Peer[1,12,13,14], Manuel Delgado-Baquerizo[15,16] & Thulani P. Makhalanyane[3,17,18]✉

Antarctic environments are dominated by microorganisms, which are vulnerable to viral infection. Although several studies have investigated the phylogenetic repertoire of bacteria and viruses in these poly-extreme environments with freezing temperatures, high ultra violet irradiation levels, low moisture availability and hyper-oligotrophy, the evolutionary mechanisms governing microbial immunity remain poorly understood. Using genome-resolved metagenomics, we test the hypothesis that Antarctic poly-extreme high-latitude microbiomes harbour diverse adaptive immune systems. Our analysis reveals the prevalence of prophages in bacterial genomes (Bacteroidota and Verrucomicrobiota), suggesting the significance of lysogenic infection strategies in Antarctic soils. Furthermore, we demonstrate the presence of diverse CRISPR-Cas arrays, including Class 1 arrays (Types I-B, I-C, and I-E), alongside systems exhibiting novel gene architecture among their effector cas genes. Notably, a Class 2 system featuring type V variants lacks CRISPR arrays, encodes Cas1 and Cas2 adaptation module genes. Phylogenetic analysis of Cas12 effector proteins hints at divergent evolutionary histories compared to classified type V effectors and indicates that TnpB is likely the ancestor of Cas12 nucleases. Our findings suggest substantial novelty in Antarctic cas sequences, likely driven by strong selective pressures. These results underscore the role of viral infection as a key evolutionary driver shaping polar microbiomes.

Keywords Adaptive immunity, Antarctica, Antiphage,, Bacteria, CRISPR-Cas, Evolutionary drivers

Understanding the ecological role played by viruses in altering ecosystem processes through their influence on both the phylogenetic and functional diversity of their hosts remains a major ecological endeavour[1–5]. In soils, viruses affect the diversity and abundance of microorganisms[6–9] thereby influencing processes such as nutrient cycling and carbon sequestration[10–13]. Their role as mediators of ecosystem services is pronounced in the poly-extreme environments of the McMurdo Dry Valleys of Eastern Antarctica where prokaryotes govern nutrient cycling. While Antarctic soils harbour diverse microbes and viruses[14–17], host-virus interactions remain

[1]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa. [2]Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology, 23955-6900 Thuwal, Saudi Arabia. [3]Department of Biochemistry, Genetics and Microbiology, Faculty of Natural and Agricultural Sciences, University of Pretoria, Hatfield, Pretoria 0028, South Africa. [4]Department of Biochemistry, Genetics and Microbiology, Faculty of Natural and Agricultural Sciences, DSI/NRF SARChI in Marine Microbiomics, University of Pretoria, Hatfield, Pretoria 0028, South Africa. [5]Scotland's Rural College, West Mains Road, Edinburgh EH9 3JG, UK. [6]School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh EH14 4AS, UK. [7]Mammal Research Institute, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa. [8]iThemba LABS, Private Bag 11, Johannesburg 2050, South Africa. [9]Canadian High Arctic Research Station, Polar Knowledge Canada; and School of Science, University of Waikato, Waitkato, New Zealand. [10]Molecular EcoSystems Biology Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, Berkeley, CA 94720, USA. [11]State Key Laboratory of Tibetan Plateau Earth System and Resources Environment, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China. [12]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium. [13]Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium. [14]Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium. [15]Laboratorio de Biodiversidad y Funcionamiento Ecosistémico, Instituto de Recursos Naturales y Agrobiología de Sevilla (IRNAS), CSIC, Seville, Spain. [16]Unidad Asociada CSIC-UPO (BioFun), Universidad Pablo de Olavide, Seville, Spain. [17]Department of Microbiology, Faculty of Science, Stellenbosch University, Stellenbosch 7600, South Africa. [18]The School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch 7600, South Africa. ✉email: tpm@sun.ac.za

relatively unexplored. This is despite the fact that viral infections may pose direct threats to microorganisms by influencing nutrient/biomass turnover, and on ecosystem functioning[18]. However, few studies have assessed these relationships, especially in poly-extreme environments, such as Antarctica, where microbial communities disproportionately influence ecosystem functions. Likewise, understanding the range of defence mechanisms used by microbes, to avoid viral infections, is crucial[19–22].

A notable prokaryotic defence mechanism, the CRISPR-Cas (clustered regularly interspaced short palindromic repeats—CRISPR associated) system[23–27], allows microorganisms to specifically target and degrade viral DNA or RNA[28–33]. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids[34,35], and have been the subject of extensive studies in the last decade owing to their efficacy in genome editing[33,36,37]. Consequently, the repertoire of known CRISPR-Cas systems has expanded significantly in terms of both quantity and diversity[35,38–41]. These systems recognize the foreign DNA of an invading phage and cleave sequences from its genome, which become integrated as spacers within the host's CRISPR array[42]. The CRISPR-Cas system thus provides signatures of previous infection events by retaining the 'genomic scars' of historical infections[43]. We hypothesise that this prokaryotic immune system may be more pronounced in poly-extreme ecosystems, where evolution is remarkably constrained and under strong selective pressures due to several persistent abiotic stressors. Because CRISPR-Cas systems represent an ancient adaptive immune strategy in prokaryotes[44] elucidating this antiviral defence pathway may reveal the immunological memories within bacterial genomes from pristine Antarctic soils. However, the extent to which these mechanisms may influence the diversity and function of terrestrial Antarctic soil communities remain poorly understood.

The Mackay Glacier region in Antarctica is one of the most remote and challenging poly-extreme environments on Earth where multiple extreme conditions including sub-zero temperatures, very low nutrient status and an absence of precipitation, render soils virtually inhospitable[45–47]. Previous, studies have revealed remarkable insights regarding the phylogenetic diversity of microbial communities in Mackay Glacier soils. For instance, these soils are predominately composed of members of the *Acidobacteriota* and *Bacteroidota* phyla[48]. These groups are known to encode a suite of antibiotic resistance genes, which may hint at possible adaptive strategies for survival in the low pH and oligotrophic conditions typical of these soils[16]. Other studies have shown that several taxa affiliated with a novel family of group 1 l [NiFe]-hydrogenases which seemingly contribute to water generation through trace gas scavenging[45]. We have also documented a diverse array of tailed bacteriophages (i.e. dsDNA phages) in these soils[14], showing that their distribution is substantially influenced by both soil pH and site altitude.

While the diversity and evolution of defence mechanisms in this ecosystem have not been studied previously, a recent studies on rock-associated Antarctic communities in the Miers Dry Valley have shown that poly-extreme environments harbour diverse defence systems[15]. The results of these studies suggest the presence of several innate immune systems including BREX (BacteRiophage EXclusion) and DISARM (Defence Island System Associated with Restriction-Modification), which were the predominant modes of antiphage immunity employed by bacteria in Antarctic desert hypoliths[15]. However, hypoliths represent only a small fraction of the Dry Valley terrain, and there is a distinct lack of studies focused on exposed surface soils.

Our overarching hypothesis is that the pristine soil microbiota of the Mackay Glacier region harbour novel adaptive immune systems. These novel systems are detectable through CRISPR-Cas arrays, which have evolved in response to the selective pressures from viral infection. Here, we aim to provide insights on adaptive immune systems through (i) characterizing the CRISPR-Cas systems within bacterial genomes, and (ii) investigating their role in viral defence mechanisms. By studying the presence of CRISPR-Cas systems in this unique ecosystem, we hope to gain a better understanding of the mechanisms by which microorganisms protect themselves from viral infections in extreme environments. We further predict that the CRISPR-Cas systems in these Antarctic soils may play crucial roles in protecting microorganisms from viral infections and maintaining the stability of the ecosystem. Using a genome-resolved metagenomic analysis, we provide the first insights of bacterial adaptive immunity, and bacterial-viral associations in Antarctic soils.

Results and discussion

Strong evidence of predation-prey associations in Antarctic soils

Our study expands current insights regarding the genetic mechanisms explaining prey-predator co-evolutionary associations between bacteria and their viruses in poly-extreme Antarctic conditions (Supplementary Table 1). Following metagenome sequencing, assembly, and genome binning (Supplementary Tables S2 and S3), we recovered 18 medium- to high-quality metagenome-assembled genomes (MAGs) (Fig. 1a). We note that this is a small sample size of reconstructed genomes, and likely reflects an underestimate of our sequencing effort (Supplementary Fig. 1). Thus, we predict that the rate of gene discovery from these sequence data would increase with higher sequencing depth. Notably, we only recovered approximately 69% of available sequence diversity from sample TG5-1, although other depth curves were approaching saturation at 12 million sequences (Table 1).

These MAGs include three *Acidobacteriota*, one *Cyanobacteria*, twelve *Bacteroidota* and two *Verrucomicrobiota*, representing both dominant, and rare bacterial phyla in these soils (Supplementary Note 1 and Supplementary Fig. 1). The genome sizes of these bacteria ranged from 2.7 – 5.9 Mb, when accounting for completeness. These genomes had moderately low G + C contents (mean = 42.8%, range 35.14%—61.44%), which is surprising given the expectation that extreme environments may select for organisms with high G + C content that allow for more stable DNA structures due to the molecular interactions of base stacking[49,50].

We estimated the genome replication rates for each MAG[51] and found that the highest genome replication rates were associated with *Acidobacteriota* (mean = 3.06) and the *Verrucomicrobiota* (mean = 2.90). These differed compared with *Bacteroidota* (mean = 2.48) and *Cyanobacteria* (2.04). Estimating the minimal doubling time with codon usage bias[52] suggested very low division times in the Antarctic bacteria (slow-growing bacteria; > 5 h), with only certain members of the *Bacteroidota* predicted to double within five hours. Specifically, the cyanobacterial

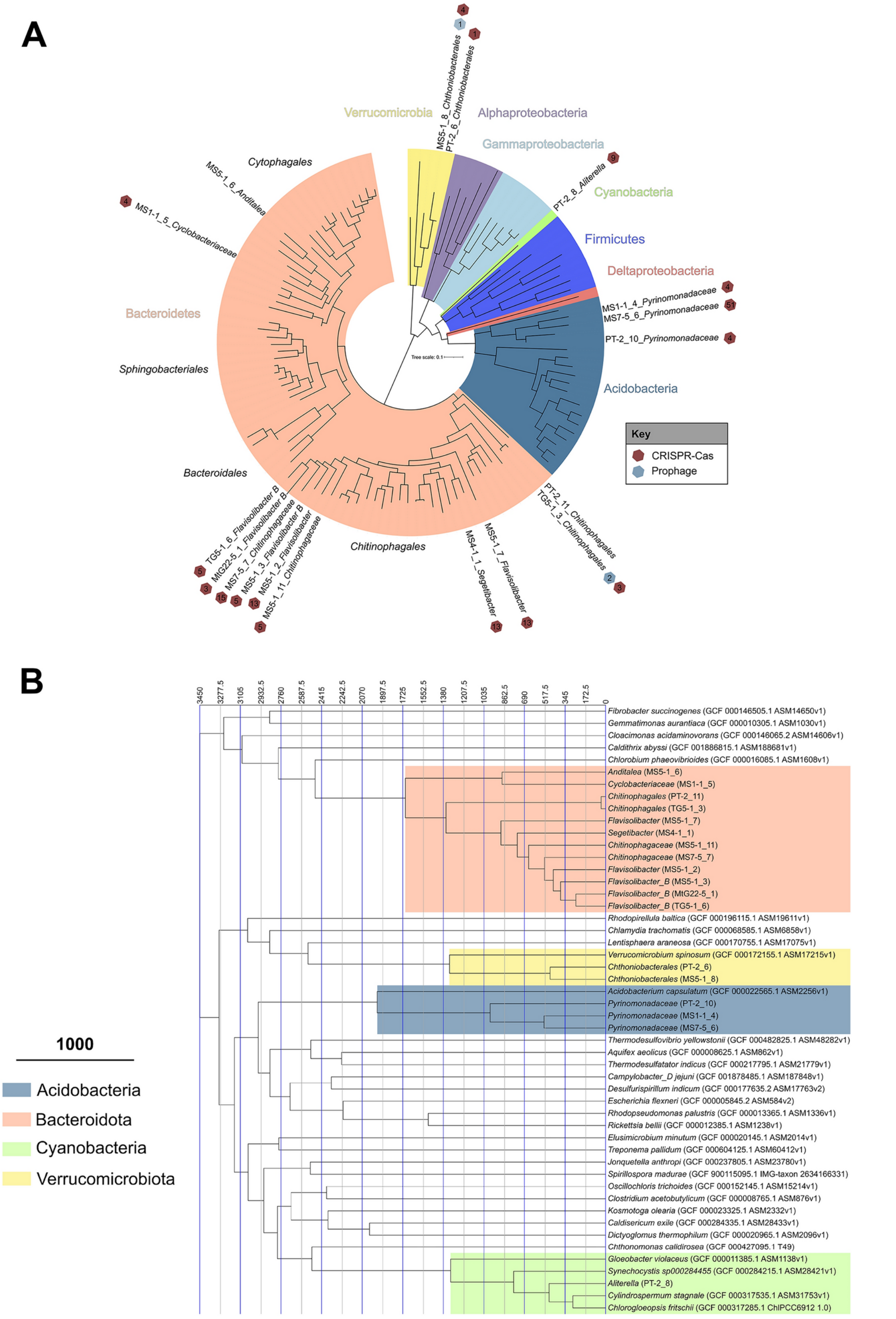
A
4
1
1
MS5-1_8_Chthoniobacterales
PT-2_6_Chthoniobacterales
Verrucomicrobia
Alphaproteobacteria
Gammaproteobacteria
9
PT-2_8_Aliterella
Cyanobacteria
Firmicutes
Deltaproteobacteria
4
MS1-1_4_Pyrinomonadaceae
51
MS7-5_6_Pyrinomonadaceae
4
PT-2_10_Pyrinomonadaceae
Acidobacteria
Cytophagales
MS5-1_6_Anditalea
4
MS1-1_5_Cyclobacteriaceae
Bacteroidetes
Sphingobacteriales
Tree scale: 0.1
Key
CRISPR-Cas
Prophage
Bacteroidales
5
TG5-1_6_Flavisolibacter B
3
MtG22-5_1_Flavisolibacter B
15
MS7-5_7_Chitinophagaceae
5
MS5-1_3_Flavisolibacter B
13
MS5-1_2_Flavisolibacter
5
MS5-1_11_Chitinophagaceae
Chitinophagales
13
MS4-1_1_Segetibacter
13
MS5-1_7_Flavisolibacter
PT-2_11_Chitinophagales
TG5-1_3_Chitinophagales
2
3
B
3450
3277.5
3105
2932.5
2760
2587.5
2415
2242.5
2070
1897.5
1725
1552.5
1380
1207.5
1035
862.5
690
517.5
345
172.5
0
Fibrobacter succinogenes (GCF 000146505.1 ASM14650v1)
Gemmatimonas aurantiaca (GCF 000010305.1 ASM1030v1)
Cloacimonas acidaminovorans (GCF 000146065.2 ASM14606v1)
Caldithrix abyssi (GCF 001886815.1 ASM188681v1)
Chlorobium phaeovibrioides (GCF 000016085.1 ASM1608v1)
Anditalea (MS5-1_6)
Cyclobacteriaceae (MS1-1_5)
Chitinophagales (PT-2_11)
Chitinophagales (TG5-1_3)
Flavisolibacter (MS5-1_7)
Segetibacter (MS4-1_1)
Chitinophagaceae (MS5-1_11)
Chitinophagaceae (MS7-5_7)
Flavisolibacter (MS5-1_2)
Flavisolibacter_B (MS5-1_3)
Flavisolibacter_B (MtG22-5_1)
Flavisolibacter_B (TG5-1_6)
Rhodopirellula baltica (GCF 000196115.1 ASM19611v1)
Chlamydia trachomatis (GCF 000068585.1 ASM6858v1)
Lentisphaera araneosa (GCF 000170755.1 ASM17075v1)
Verrucomicrobium spinosum (GCF 000172155.1 ASM17215v1)
Chthoniobacterales (PT-2_6)
Chthoniobacterales (MS5-1_8)
Acidobacterium capsulatum (GCF 000022565.1 ASM2256v1)
Pyrinomonadaceae (PT-2_10)
Pyrinomonadaceae (MS1-1_4)
Pyrinomonadaceae (MS7-5_6)
Thermodesulfovibrio yellowstonii (GCF 000482825.1 ASM48282v1)
Aquifex aeolicus (GCF 000008625.1 ASM862v1)
Thermodesulfatator indicus (GCF 000217795.1 ASM21779v1)
Campylobacter_D jejuni (GCF 001878485.1 ASM187848v1)
Desulfurispirillum indicum (GCF 000177635.2 ASM17763v2)
Escherichia flexneri (GCF 000005845.2 ASM584v2)
Rhodopseudomonas palustris (GCF 000013365.1 ASM1336v1)
Rickettsia bellii (GCF 000012385.1 ASM1238v1)
Elusimicrobium minutum (GCF 000020145.1 ASM2014v1)
Treponema pallidum (GCF 000604125.1 ASM60412v1)
Jonquetella anthropi (GCF 000237805.1 ASM23780v1)
Spirillospora madurae (GCF 900115095.1 IMG-taxon 2634166331)
Oscillochloris trichoides (GCF 000152145.1 ASM15214v1)
Clostridium acetobutylicum (GCF 000008765.1 ASM876v1)
Kosmotoga olearia (GCF 000023325.1 ASM2332v1)
Caldisericum exile (GCF 000284335.1 ASM28433v1)
Dictyoglomus thermophilum (GCF 000020965.1 ASM2096v1)
Chthonomonas calidirosea (GCF 000427095.1 T49)
Gloeobacter violaceus (GCF 000011385.1 ASM1138v1)
Synechocystis sp000284455 (GCF 000284215.1 ASM28421v1)
Aliterella (PT-2_8)
Cylindrospermum stagnale (GCF 000317535.1 ASM31753v1)
Chlorogloeopsis fritschii (GCF 000317285.1 ChlPCC6912 1.0)
1000
Acidobacteria
Bacteroidota
Cyanobacteria
Verrucomicrobiota

Fig. 1. (**A**) Maximum likelihood tree of all metagenome-assembled genomes (MAGs). Phylogenetic tree constructed with a concatenated alignment of 49 core, bacterial genes. The tree includes 18 bacterial MAGs from this study that are aligned to 100 known reference genomes present in the RefSeq database. Hexagons adjacent to the genome names indicate either the presence of a CRISPR-Cas repeat (maroon) or a prophage (grey) within the host genome. Numbers in hexagons indicate counts of prophages or CRISPR spacers. (**B**) Bayesian divergence estimates of our Antarctic MAGs placed among 32 reference outgroup genomes. Time scale is estimated in mega-annum.

Bin Id	Lineage (GTDB-Tk)	Genome Size (bp)	Completeness (%)	Contamination (%)	Prophages	CRISPR spacers
MS1-1.4	*Pyrinomonadaceae*	32,91,545	82.38	6.46	0	4
MS1-1.5	*Cyclobacteriaceae*	33,19,274	75.94	4.27	0	4
MS4-1.1	*Segetibacter*	26,59,359	58.37	3.45	0	13
MS5-1.11	*Chitinophagaceae*	38,56,181	94.88	2.96	0	5
MS5-1.2	*Flavisolibacter*	33,35,429	97.49	2.96	0	13
MS5-1.3	*Flavisolibacter_B*	16,54,061	73.45	1.72	0	5
MS5-1.6	*Anditalea*	30,47,815	51.02	1.72	0	0
MS5-1.7	*Flavisolibacter*	31,49,465	78.72	1.23	0	13
MS5-1.8	*Chthoniobacterales*	26,52,899	66.01	6.45	1	4
MS7-5.6	*Pyrinomonadaceae*	39,55,314	91.10	9.47	0	51
MS7-5.7	*Chitinophagaceae*	34,32,039	91.89	9.28	0	15
MtG22-5.1	*Flavisolibacter_B*	28,54,497	50.86	3.45	0	3
PT-2.10	*Pyrinomonadaceae*	30,82,920	53.69	3.59	0	4
PT-2.11	*Chitinophagales*	25,44,403	57.68	4.16	0	0
PT-2.6	*Chthoniobacterales*	24,95,165	89.53	5.17	0	1
PT-2.8	*Aliterella*	23,33,589	64.44	3.45	0	9
TG5-1.3	*Chitinophagales*	16,45,326	83.62	6.45	2	3
TG5-1.6	*Flavisolibacter_B*	26,76,035	82.79	6.16	0	5

Table 1. Genomic statistics for the 18 MAGs recovered from the Mackay Glacier region of Antarctica.

genome (> 21 h) and the Verrucomicrobiota genomes (> 15 h) indicated very low division times. Overall, these patterns suggest extremely slow growth rates which is expected given the poly-extreme conditions in this region (see additional information regarding the climatic conditions and these taxa in Supplementary Note 1). Given the low division rates, we believe that it becomes reasonable to predict that the slow evolutionary processes acting on microbial communities in this environment are likely to delay selection. There is some support for this assertion including the divergent ecological patterns of Antarctic soil microbiota[53,54] and the substantial differences in community composition compared with soils from outside the continent[45].

Our study provides strong evidence that Antarctic bacterial communities may have ancient origins. Bayesian evolutionary analysis, used to produce time-measured phylogenies, suggests that the genomes retrieved from our studies ranged between 500 and 1,200 Mya (Fig. 1b). These findings are consistent with previous estimates of cryptoendolith divergence in Antarctica[55]. The results also support approximations for other genomes retrieved from Antarctic soils[45]. Altogether, the unique monophyletic clades of our Antarctic MAGs were distinct—suggesting that these bacteria diverged from other known taxa during the Precambrian (541 Mya). Taxonomic analysis indicated that 12 of our MAGs potentially represent novel species as they show low homologies to those available in reference databases. Considering this evidence, we predict that bacteriophages associated with these bacteria may have also been co-separated from other microorganisms for a similar length of time since host-virus specificity is mostly strain specific. We hypothesized that this distinct, and specific, co-evolution may be corroborated by the recovery of uncharacterized and potentially novel adaptive immune signatures in Antarctic host genomes (see Supplementary Note 2 on viral sequence analyses).

CRISPR-Cas systems provide evidence of virus attacks in Antarctica

The detection of prophages in MAGs and AMGs on phage contigs (Supplementary Note 2) supports the notion of unique host-virus histories, as almost all our Antarctic viral genomes are completely unrelated to known viruses based on protein similarities (Fig. 2), as observed in many unexplored habitats previously[56–58]. The results from this study suggest that the host adaptive immune system, associated with divergent microbiota in Antarctica soils, may be more prominent than initially envisaged. However, apart from previous studies on rock associated microbiota, there is a severe knowledge deficit regarding host adaptive immune systems associated with poly-extreme environments.

To further explore host-virus histories associated with our MAGs, we searched for related diverse defence strategies against phage predation. As part of determining the adaptive immune system, we found putative CRISPR-Cas systems in 16 of the 18 MAGs (*ca.* 89%). In terms of CRISPR arrays, the identified conserved repeats ranged from 23 to 30 bp across four of the retrieved genomes. The repeats were flanked by unique spacers (obtained from viruses), that were an average of 36 bp in length (range 34—38 bp). The largest set of CRISPR cassettes was found in the MS7-5_6 genome (*Acidobacteriota*), with 51 CRISPR unique spacers housed between 6 different CRISPR cassettes. These values are within the optimal number of spacers, previously suggested to range between 10 – 100 spacers within bacterial genomes[59]. The CRISPR loci within bacterial genomes retain the memory of past viral infections and foreign DNA encounters[59,60], suggesting at least 50 previous such events. Yet, the length of these loci appears to be directly related to the capacity to respond to an infection[61]. In other words, there appears to be a trade-off between maintaining a vast genetic memory of attacks (harbouring more spacers) and the functionality of the CRISPR mechanism[59]. One hypothesis for this is that array size represents a balance between maintaining immunity to many potential threats (older infections) and updating immunity

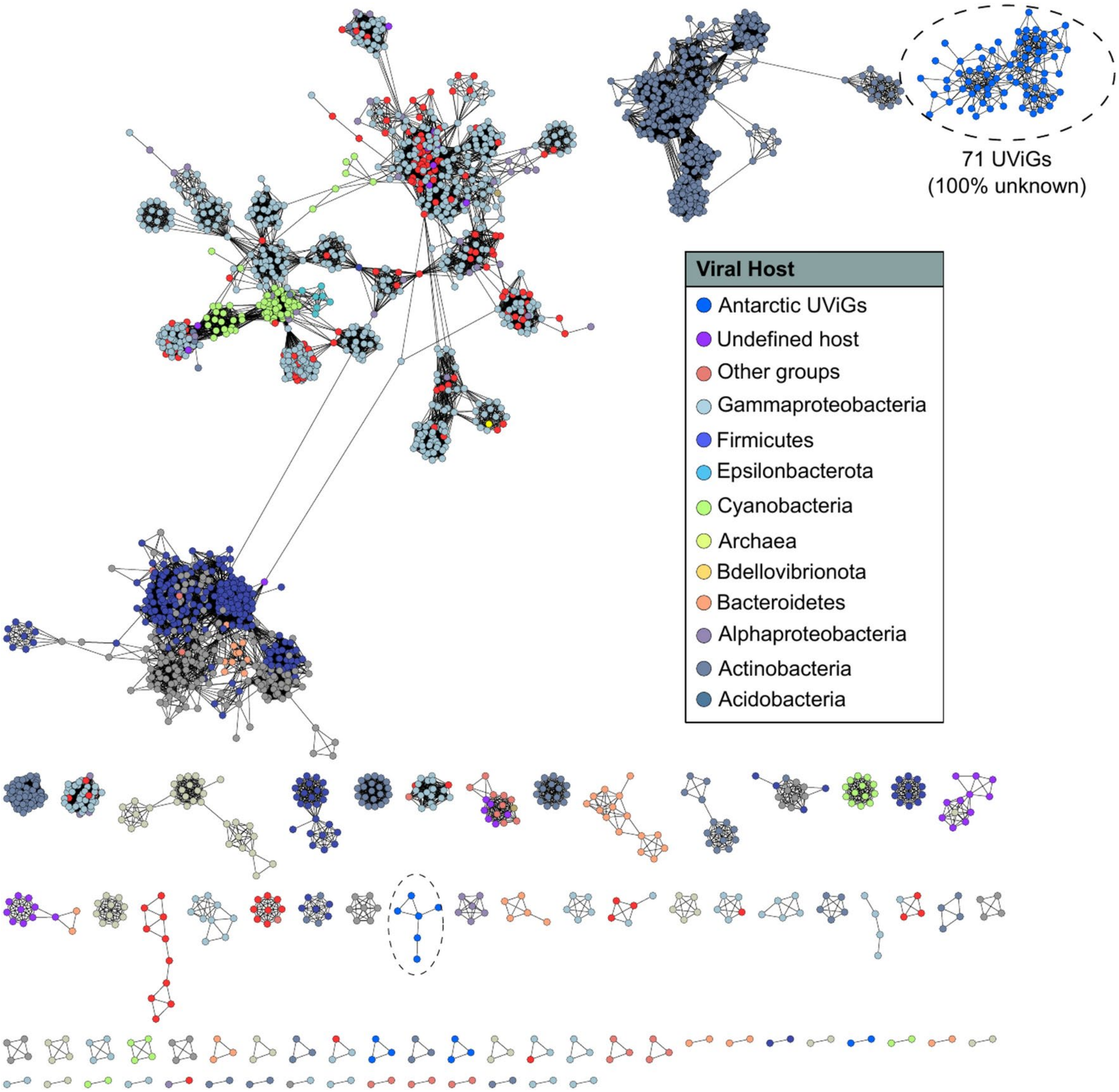

Fig. 2. Viral contigs in MAGs and metagenomic datasets. Viral protein cluster network (PC) produced using Cytoscape v3.8.2. Network showing the clustering of recovered viral genomes and viral genome fragments with known RefSeq viral genomes (v85) based on their shared predicted protein content. Major viral families (nodes) are indicated by the colour of their host, with vOTUs found in this study indicated by blue circles and with dotted lines. The edges (lines between nodes) indicate significant shared protein content.

to contend with new threats[62]. The remaining genomes only had between three and 16 spacers, which is more similar to human gut microbes (average of 12 spacers)[63] than the average cassette size of between 20 and 40 spacers[64]. We speculate that the lower spacer count may be due to limited encounters with a small set of phages. In this scenario, the spatial constraints of the soil microhabitat limit the number of potential interactions between phages and putative hosts. This suggests that phage diversity may be low in this region of Antarctica. Not only are cells immobilized by adsorption to soil particles of the Antarctic desert pavement, but rarely, if ever, subject to precipitation events which may allow for the mobilization of cells or virus-like particles, thus reducing the spectrum of infection events considerably.

In addition to the CRISPR-Cas cassettes, 16 MAGs had relatively low abundances of *cas* genes, with between 6 and 42 loci per MAG. These *cas* genes constituted 122 unclassified sequences ($n = 221$ total *cas* sequences), followed by several classified sequences including 48 type III, 31 type I, 20 type IV and 2 type V Cas systems. These Cas types are similar to those previously reported in Antarctic surface snow in which CRISPR-cas types I, II and III were most common[65]. The MS7-5_6 MAG (*Acidobacteriota*) had a contig with 10 genes associated with a hybrid CRISPR-Cas Class I system. This contig also had a GCN5-related N-acetyltransferase (GNAT) toxin domain[66] (see Fig. 3a), which functions by acetylating charged tRNA molecules to prevent translation. Previous studies suggest that these GCN5-related N-acetyltransferase toxin domains may represent novel substrates for several enzymes linked to antibiotic modification[67].

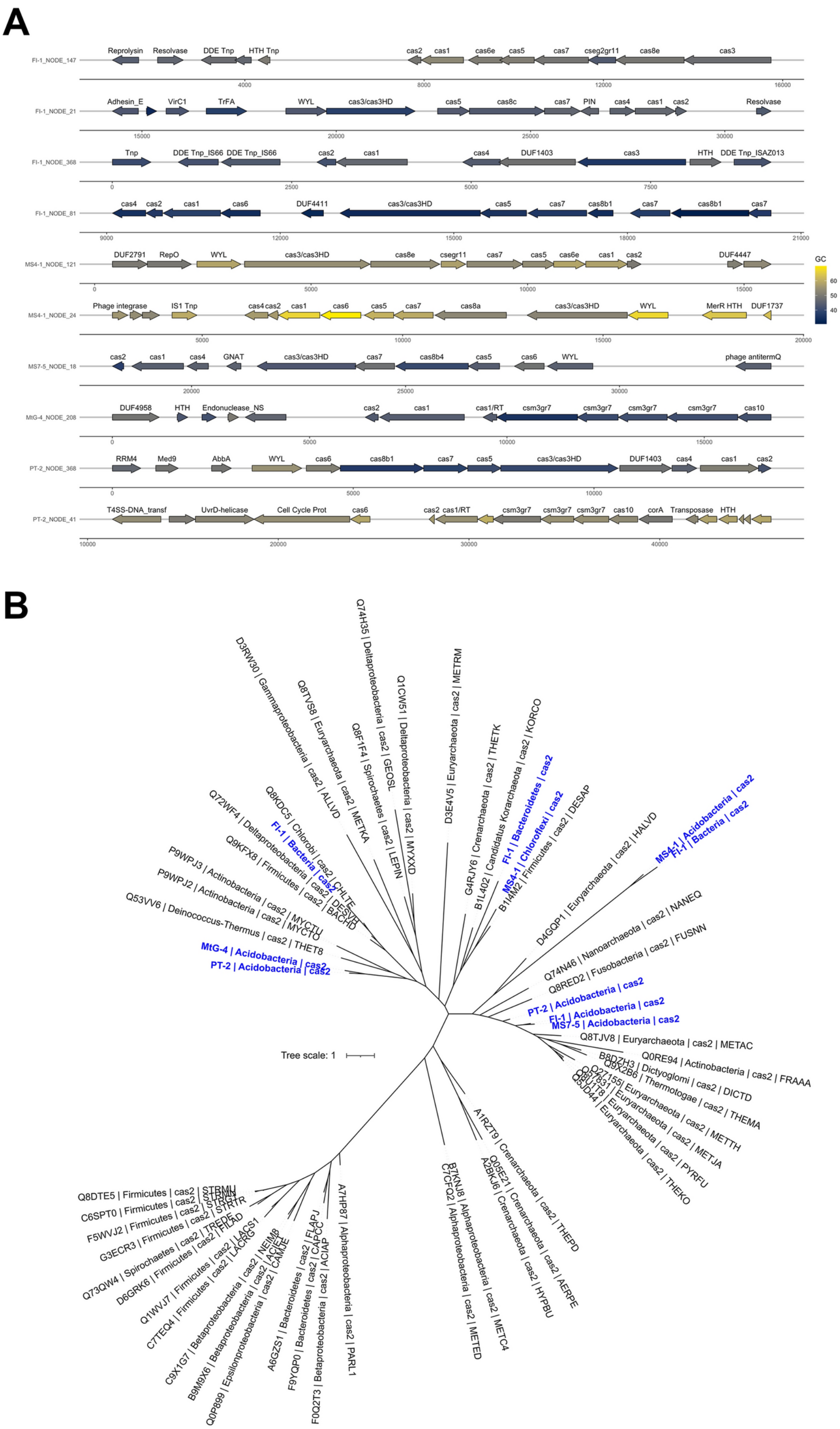
A
FI-1_NODE_147
Reprolysin
Resolvase
DDE Tnp
HTH Tnp
cas2
cas1
cas6e
cas5
cas7
cseg2gr11
cas8e
cas3
4000
8000
12000
16000
FI-1_NODE_21
Adhesin_E
VirC1
TrFA
WYL
cas3/cas3HD
cas5
cas8c
cas7
PIN
cas4
cas1
cas2
Resolvase
15000
20000
25000
30000
FI-1_NODE_368
Tnp
DDE Tnp_IS66
DDE Tnp_IS66
cas2
cas1
cas4
DUF1403
cas3
HTH
DDE Tnp_ISAZ013
0
2500
5000
7500
FI-1_NODE_81
cas4
cas2
cas1
cas6
DUF4411
cas3/cas3HD
cas5
cas7
cas8b1
cas7
cas8b1
cas7
9000
12000
15000
18000
21000
MS4-1_NODE_121
DUF2791
RepO
WYL
cas3/cas3HD
cas8e
csegr11
cas7
cas5
cas6e
cas1
cas2
DUF4447
0
5000
10000
15000
GC
60
50
40
MS4-1_NODE_24
Phage integrase
IS1 Tnp
cas4
cas2
cas1
cas6
cas5
cas7
cas8a
cas3/cas3HD
WYL
MerR HTH
DUF1737
5000
10000
15000
20000
MS7-5_NODE_18
cas2
cas1
cas4
GNAT
cas3/cas3HD
cas7
cas8b4
cas5
cas6
WYL
phage antitermQ
20000
25000
30000
MtG-4_NODE_208
DUF4958
HTH
Endonuclease_NS
cas2
cas1
cas1/RT
csm3gr7
csm3gr7
csm3gr7
csm3gr7
cas10
0
5000
10000
15000
PT-2_NODE_368
RRM4
Med9
AbbA
WYL
cas6
cas8b1
cas7
cas5
cas3/cas3HD
DUF1403
cas4
cas1
cas2
0
5000
10000
PT-2_NODE_41
T4SS-DNA_transf
UvrD-helicase
Cell Cycle Prot
cas6
cas2
cas1/RT
csm3gr7
csm3gr7
csm3gr7
cas10
corA
Transposase
HTH
10000
20000
30000
40000
B
D3RW30 | Gammaproteobacteria | cas2 | ALLVD
Q8TVS8 | Euryarchaeota | cas2 | METKA
Q74H35 | Deltaproteobacteria | cas2 | GEOSL
Q8F1F4 | Spirochaetes | cas2 | LEPIN
Q1CW51 | Deltaproteobacteria | cas2 | MYXXD
D3E4V5 | Euryarchaeota | cas2 | METRM
G4RJY6 | Crenarchaeota | cas2 | THETK
B1L402 | Candidatus Korarchaeota | cas2 | KORCO
FI-1 | Bacteroidetes | cas2
MS4-1 | Chloroflexi | cas2
B1I4M2 | Firmicutes | cas2 | DESAP
D4GQP1 | Euryarchaeota | cas2 | HALVD
MS4-1 | Acidobacteria | cas2
FI-1 | Bacteria | cas2
Q74N46 | Nanoarchaeota | cas2 | NANEQ
Q8RED2 | Fusobacteria | cas2 | FUSNN
PT-2 | Acidobacteria | cas2
FI-1 | Acidobacteria | cas2
MS7-5 | Acidobacteria | cas2
Q8TJV8 | Euryarchaeota | cas2 | METAC
Q0RE94 | Actinobacteria | cas2 | FRAAA
B8DZH3 | Dictyoglomi | cas2 | DICTD
Q9X2B6 | Thermotogae | cas2 | THEMA
Q27155 | Euryarchaeota | cas2 | METTH
Q57831 | Euryarchaeota | cas2 | METJA
Q8U1T8 | Euryarchaeota | cas2 | PYRFU
Q5JD44 | Euryarchaeota | cas2 | THEKO
A1RZT9 | Crenarchaeota | cas2 | THEPD
Q05E21 | Crenarchaeota | cas2 | AERPE
A2BKJ6 | Crenarchaeota | cas2 | HYPBU
B7KNJ8 | Alphaproteobacteria | cas2 | METC4
C7CFQ2 | Alphaproteobacteria | cas2 | METED
A7HP87 | Alphaproteobacteria | cas2 | PARL1
F0Q2T3 | Betaproteobacteria | cas2 | ACIAP
F9YQP0 | Bacteroidetes | cas2 | CAPCC
A6GZS1 | Bacteroidetes | cas2 | FLAPJ
Q0P899 | Epsilonproteobacteria | cas2 | CAMJE
B9M9X6 | Betaproteobacteria | cas2 | ACIET
C9X1G7 | Betaproteobacteria | cas2 | NEIM8
C7TEQ4 | Firmicutes | cas2 | LACRG
Q1WVJ7 | Firmicutes | cas2 | LACS1
D6GRK6 | Firmicutes | cas2 | FILAD
Q73QW4 | Spirochaetes | cas2 | TREDE
G3ECR3 | Firmicutes | cas2 | STRTR
F5WVJ2 | Firmicutes | cas2 | STRGT
C6SPT0 | Firmicutes | cas2 | STRMN
Q8DTE5 | Firmicutes | cas2 | STRMU
Tree scale: 1
Q53VV6 | Deinococcus-Thermus | cas2 | THET8
MtG-4 | Acidobacteria | cas2
PT-2 | Acidobacteria | cas2
P9WPJ2 | Actinobacteria | cas2 | MYCTO
P9WPJ3 | Actinobacteria | cas2 | MYCTU
Q9KFX8 | Firmicutes | cas2 | BACHD
Q72WF4 | Deltaproteobacteria | cas2 | DESVH
FI-1 | Bacteria | cas2
Q8KDC5 | Chlorobi | cas2 | CHLTE

Fig. 3. Cas proteins revealed varied taxonomic histories. (**A**) The contigs containing Cas genes are shown. They are colored by their G + C content which showed substantial variation across the contigs. (**B**) Phylogenetic tree of Cas2 genes recovered from our metagenomes (shown in blue) with reference sequences in black. Blue labels indicate both the sampling site and taxonomy. Tree scale is shown on the left.

We further investigated unbinned metagenomic contigs, which possessed eight or more co-localized *cas* genes, to determine if they represented novel CRISPR-Cas variants. Taxonomically, the CRISPR-Cas systems recovered from these contigs were affiliated members of the *Acidobacteriota* ($n = 6$ contigs), Unclassified Bacteria ($n = 2$), *Chloroflexota* ($n = 1$) and *Bacteroidota* ($n = 1$). However, the taxonomic relationships of these taxa suggest potentially shared histories with a variety of bacterial phyla (Fig. 3b). The architecture of effector complexes, within the CRISPR-Cas systems, suggests that most of these were class 1 with type I or type III systems. Genes for Cas1 and Cas2 proteins were ubiquitously distributed across all contigs (Fig. 3a). These genes were always structured as Cas1-Cas2 complexes. In four examples, the Cas1-Cas2 complex was flanked upstream by *cas*4, which directly interacts with the Cas1-Cas2 complex, to process pre-spacers prior to integration as the Cas4-Cas1-Cas2 complex[68]. However, in two instances, our analyses revealed that *cas*4 was downstream of the Cas1-Cas2 complex, which is an unconventional arrangement of these genes based on data from previous studies[33]. In all 10 cases, the effector genes were located upstream of the Cas1-Cas2 operon.

The remaining four CRISPR-Cas systems may represent novel variants, based on arrangements of their effector modules (Fig. 3b)[33,69]. These results imply ongoing horizontal gene transfer and recombination events of diverse CRISPR-Cas loci, probably led by continuous interactions with the same viruses. Notably, these uncategorized CRISPR-Cas system types were affiliated with members of the *Acidobacteriota*. They include FI-1_NODE_368 (cas2-cas1-cas4-cas3), which lacks an effector complex and seems to be closely related to Type IU. Contig FI-1_NODE_81 (cas4-cas2-cas1-cas6-cas3-cas5-cas7-cas8b1-cas7-cas8b1-cas7) a potential Type I-B variant based on multiple copies of cas7 and cas8 at the terminus of the array. Contig MtG-4_NODE_208 (cas2-cas1-cas1-RT-csm3gr7-csm3gr7-csm3gr7-cas10), potentially a Type IIIU array with three copies of csm3gr7, and finally, contig PT-2_NODE_41 (cas6-cas2-cas1-csm3gr7-csm3gr7-csm3gr7-cas10) which may be a Type IIIA variant or Type IIIU variant since it lacks csm2, csm4 and csm5 genes.

All 10 predicted CRISPR-Cas systems were associated with CRISPR arrays. These systems were composed of spacers that ranged from 2 to 122 bp in length, with an average length of 35 bp. The *cas*2 sequences showed some divergence from those previously reported, and these results contrasted with our expectations. Instead, the *cas*2 sequences clustered among unrelated phyla, in some cases grouping within the domain *Archaea* (Fig. 3a). These results are not surprising given the fact that these genes are known to be horizontally acquired. This may indicate that the *cas*2 gene is not always taxonomically conserved. Instead, the result suggests mobilization via inter-phylum horizontal gene transfer (HGT) events or evidence of phylum-specific *cas* subtypes. A recent study showed that CRISPR-Cas systems may contribute to the propagation of transposable elements by facilitating transposition into specific sites[70]. Similarly, our results support previous reports since we found transposase elements on almost half of the 10 CRISPR-Cas-containing contigs analysed.

Based on these data, we speculated that these Antarctic CRISPR-Cas systems were horizontally transferred as ancient mobilization events. This suggestion is supported by an evaluation of the G + C skew, among the 10 contigs containing *cas* genes, as a proxy for the timing of insertion events[71]. Here, we inferred HGT through the detection of strong deviations in G + C content for a genomic fragment compared to the remaining genomic signature. Specifically, on NODE_81 from the FI-1 metagenome, the G + C content over the Cas proteins varied minimally across each gene yet is markedly different from the G + C content of the CRISPR array upstream of the *cas* genes (Fig. 4). By contrast, the contigs containing integrated prophages within the microbial genomes showed very high variations in G + C content (i.e. G + C skew) across the contig which hints at their foreign origin[72] (Supplementary Fig. 3). Our Bayesian diversity estimates also indicated ancient divergence events of our MAGs from known bacteria. It is thus likely that the phages of these bacteria have similarly ancient Precambrian histories, which offers a possible explanation for their unique gene compositions.

Following this, we explored our data for the diversity of type V CRISPR-Cas systems. From the data, we identified a total of 216 contigs longer than 1 kb from 16 of the 18 metagenomes with predicted cas12 effectors proteins. Of these, 112 contigs with sizes ranging from 1,007 to 48,306 bp that possessed non-partial cas12 proteins were retained for downstream analyses. The lengths of effector proteins in these contigs varied from 89 to 630 amino acids, and this contrasted with previous reports that have indicated the average lengths for type V associated effector proteins to be ~ 400 amino acids and longer[73,74]. As effector proteins associated with type V CRISPR-Cas systems are mainly distinguished by the possession of a RuvC nuclease domain, we also found these to be characteristic of our 111 effectors, including the smallest (89 aa) putative cas12 protein. Only one of these lacked a RuvC domain but instead possessed a rudimentary helix-turn-helix domain. Further inspection of contigs possessing these indicated that only 13 of our effectors were proximal to CRISPR arrays, and unlike typical CRISPR-Cas systems none of the 112 were co-localized with the cas1-cas2 complex. Phylogenetic analysis of these indicated that just nine of our effectors (Ant Cas U5-8) clustered with previously characterized cas12 effectors. We then observed that 18 of our other effectors indicated a close phylogenetic relationship with transposon encoded TnpB proteins, suggesting that the Antarctic type V effectors may have evolved from TnpB associated nucleases, which has been speculated previously[75]. We observed a further 83 additional effectors from our data that formed a distinct clade (indicated as Ant Cas U4), potentially representing a novel subgroup of cas12-like effectors (Fig. 5).

Altogether, we speculate that the unique diversity of the genes found in these Antarctic soils may be the result of a 'slowed down' evolution of genes selected during warmer periods of time. The Antarctic continent was a temperate rainforest during the mid-Cretaceous period ~ 140 Mya[76] and we speculate that the subsequent cooling of the continent may have constrained evolutionary forces from acting at their previous pace. Combined, these lines of evidence point to an ancient, acquired immunity of bacteria in Antarctica while contemporary infection events continue to occur through lysogenic phage infections.

Conclusions

We used metagenomes from remote and pristine Antarctic soils to assess their viral and bacterial diversity. Multiple lines of evidence suggest extensive phage-host interactions, potentially novel viral diversity, and CRISPR-Cas variants. The phage signatures (vOTUs) were linked to the infection of dominant soil bacterial lineages in these surface soils, including members of the *Bacteroidota* and *Acidobacteriota*, while prophages embedded within *Verrucomicrobiota* and *Bacteroidota* MAGs offer further insight into contemporary infections. CRISPR-

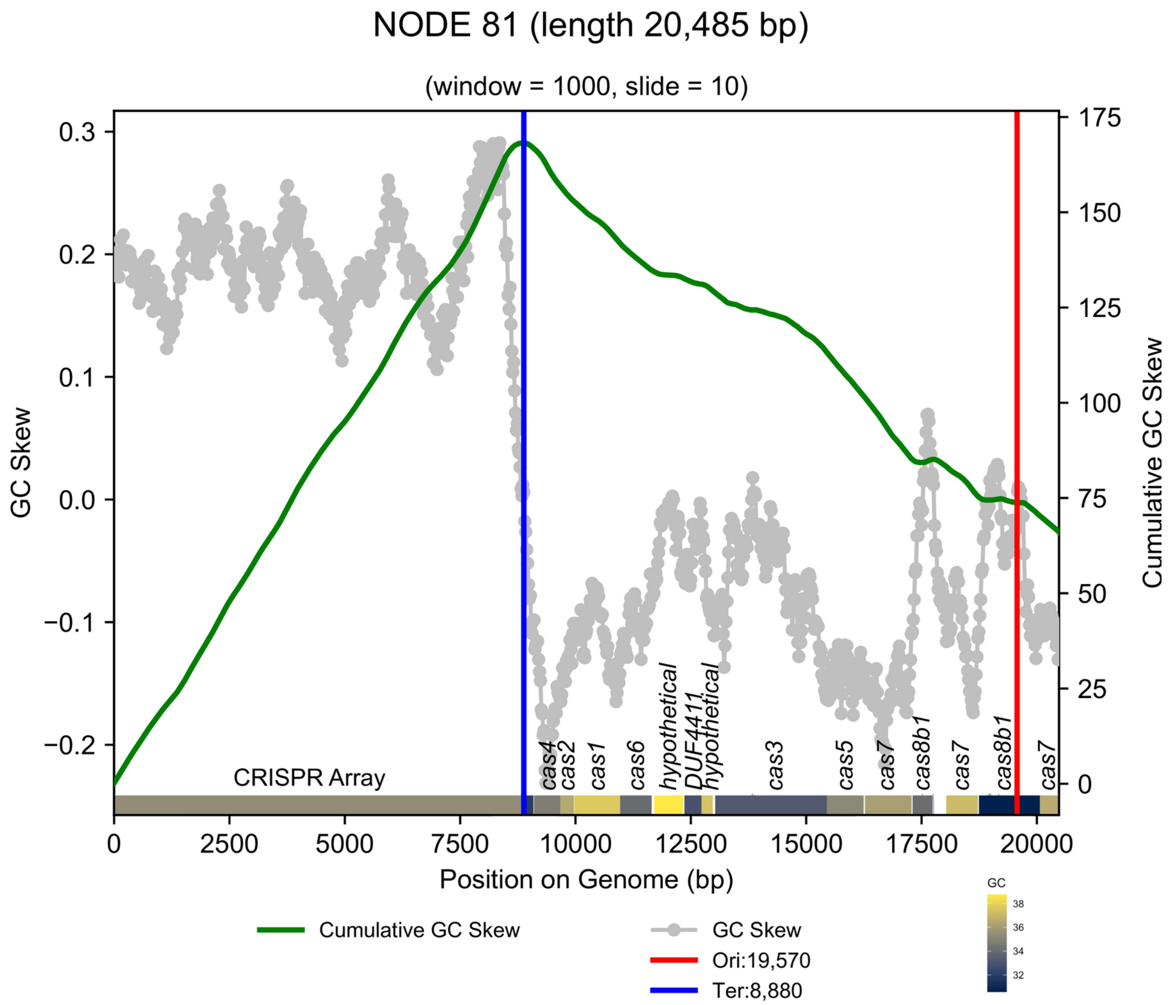

Fig. 4. Plot indicating GC skew across contigs. G + C skew across NODE_81 is shown along the contig (x-axis plots gene position in the contig) with G + C skew on the left-hand y-axis (grey points) and Cumulative G + C skew on the right-hand y-axis (green line). The predicted origin and termination of replication are shown in red and blue lines, respectively. The CRISPR-Cas array is shown at the bottom of the figure with the G + C content of each gene shown.

Cas systems, part of the bacterial adaptive immune system, were common to 4 of 18 MAGs analyzed, indicating acquired immunity in both *Bacteroidota* and *Acidobacteriota*. Additional Class I CRISPR-Cas arrays (types I-B, I-C and I-E) were detected in the assembled metagenomes, where four CRISPR-Cas arrays did not match existing architectures and thus may be novel variants. The difference in the architecture observed within our CRISPR-Cas arrays suggests that these may provide an enhanced ability for bacteria to resist viral predation and adsorption. This may be a result of extreme conditions in the Antarctic which may influence viruses to employ different infection strategies, and expand on the diversity and versatility of CRISPR-Cas systems[77]. Our analysis of G + C content and GC skew across CRISPR-Cas contigs showed low variations in G + C skew in CRISPR-Cas arrays, but more variation in prophages, suggesting that these acquired immunity markers are ancient whereas proviral elements appear to be the result of recent foreign DNA transfer. This is further evidenced by the description of novel, Antarctic-exclusive cas12-like effectors with remarkable sequence homology to TnpB transposases, suggesting that TnpB could be an ancestor of Cas12 nucleases adopted by CRISPR-Cas systems.

Materials and Methods

Sample collection and preparation

Surface soils were collected from 18 remote sites in Eastern Antarctica, between the Mackay Glacier (76.52°S 161.45°E) and the Drygalski Ice Tongue (Fig. 6)[16]. Methods of DNA isolations, soil physicochemical analyses, soil isotopic measurements and soil respiration experiments have been reported previously[48]. Briefly, the Mackay Glacier—located to the north of the McMurdo Dry Valleys, Victoria Land, Antarctica—were sampled for surface mineral soil samples from 18 ice-free sites. At each of the 18 sites, approximately 20 g soil samples were retrieved aseptically from five positions within a 1 m^2 quadrat (0–5 cm soil depth), providing 90 soil samples in all. Soils were stored in sterile Whirl Pak bags (Nasco International, WI, USA) on ice during sampling and transport in the Antarctic, and at—80 °C in the laboratory (Centre for Microbial Ecology and Genomics, University of Pretoria,

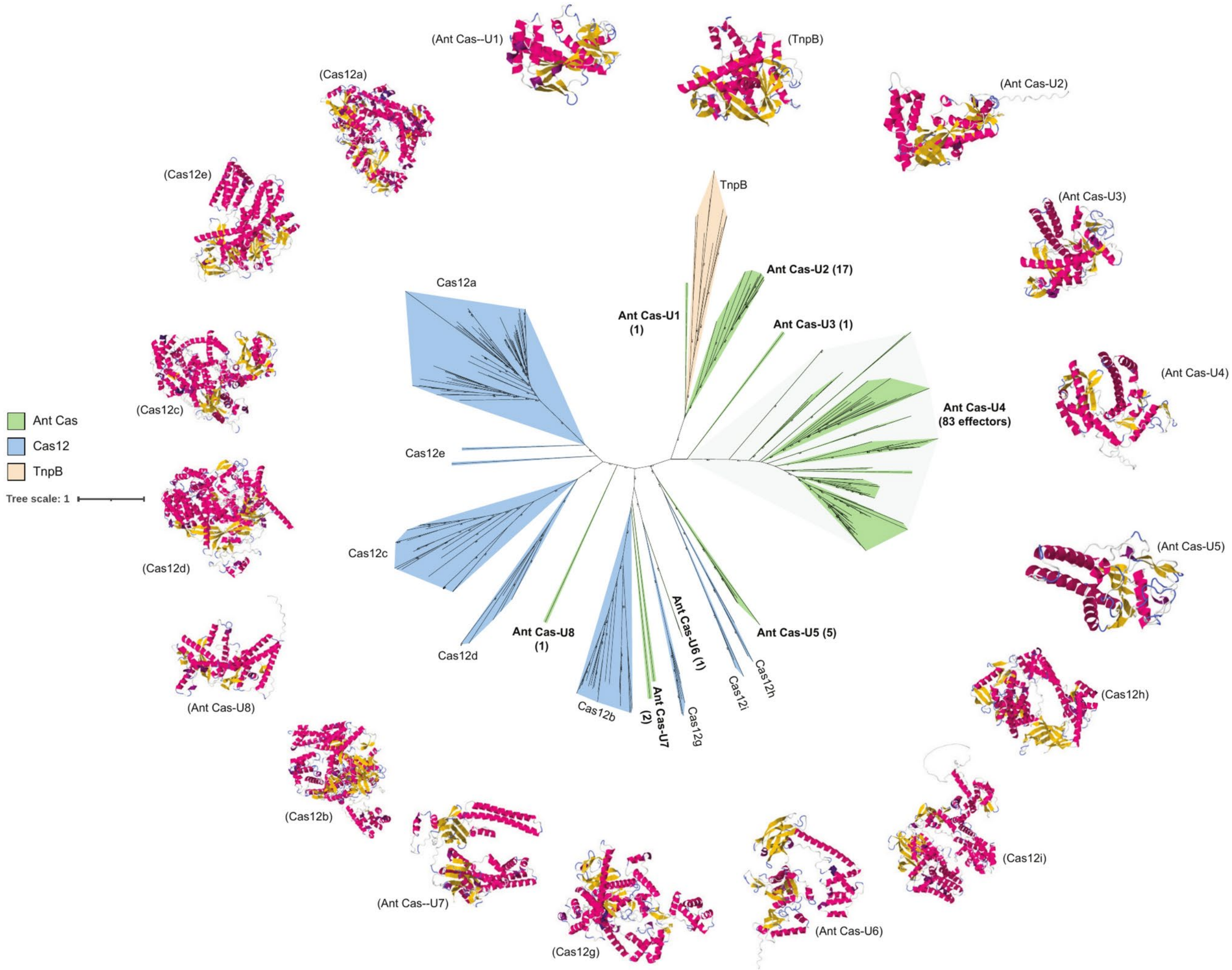

Fig. 5. A novel subgroup of cas12-like effectors identified from pristine Antarctic soils. Novel Antarctic cas12 effectors (Ant Cas-U1—Ant Cas-U8; shown in green) show unique phylogenies from known cas12 effectors (Cas12a-Cas12i; blue). The clades Ant Cas-U1 and Ant Cas-U2 show a shared evolutionary history with TnpB nucleases (orange) and have similar predicted tertiary protein structures (outer ring).

South Africa) until processed. After DNA extractions, metagenomes were sequenced on an Illumina HiSeq 2000 instrument producing 250 bp paired-end libraries. Sequencing depth was determined using Nonpareil v3.301[78].

Metagenome analysis and assembly

All 18 metagenomes were filtered to cull low-quality reads and those containing ambiguous bases (internal *N*'s) using Prinseq-lite v0.20.4[79]. Further filtering parameters included removing the sequencing adapters, reads with Phred scores < 20 for six consecutive bases, unpaired reads, and reads shorter than 100 bp in length. All Illumina PhiX sequences were identified and removed using BBDuk[80] to eliminate the potential of contaminating viral signals in our analysis[81]. We determined the microbial taxonomy of each sample using SingleM, which relies on analyses of universal single-copy ribosomal subunit proteins (*rplB*), rather than the 16S rRNA gene to infer taxonomy (https://github.com/wwood/singlem). Each filtered metagenome was individually assembled using metaSPAdes v3.12[82] under default settings with *k*-mer step increases from 21 to 141. We used MicrobeCensus v1.1.0 to calculate the number of genome equivalents in each sample[83].

Reconstruction of microbial genomes

From the 18 assemblies (one from each site), we reconstructed microbial genome bins using MetaBAT 2 v2.12.1[84]. All contigs > 1,500 bp in length were retained and depth coverage information was obtained using BBMap[80] by mapping corresponding metagenomic reads back to those contigs. The bins were assessed for completeness using CheckM2 v1.0.2[85] and metagenome-assembled genomes (MAGs) that were > 50% complete and were < 10% contaminated were retained for further analyses. Next, we queried indicators of genome quality, such as the presence of 5S, 16S and 23S ribosomal subunit genes and the presence of at least 18 unique tRNAs according to MIMAG standards[86]. Taxonomy was then assigned using the Genome Taxonomy Database Toolkit (GTDB-Tk) v2.3.2 release version 214[87]. CARD RGI was used to determine the prevalence of antibiotic resistance genes

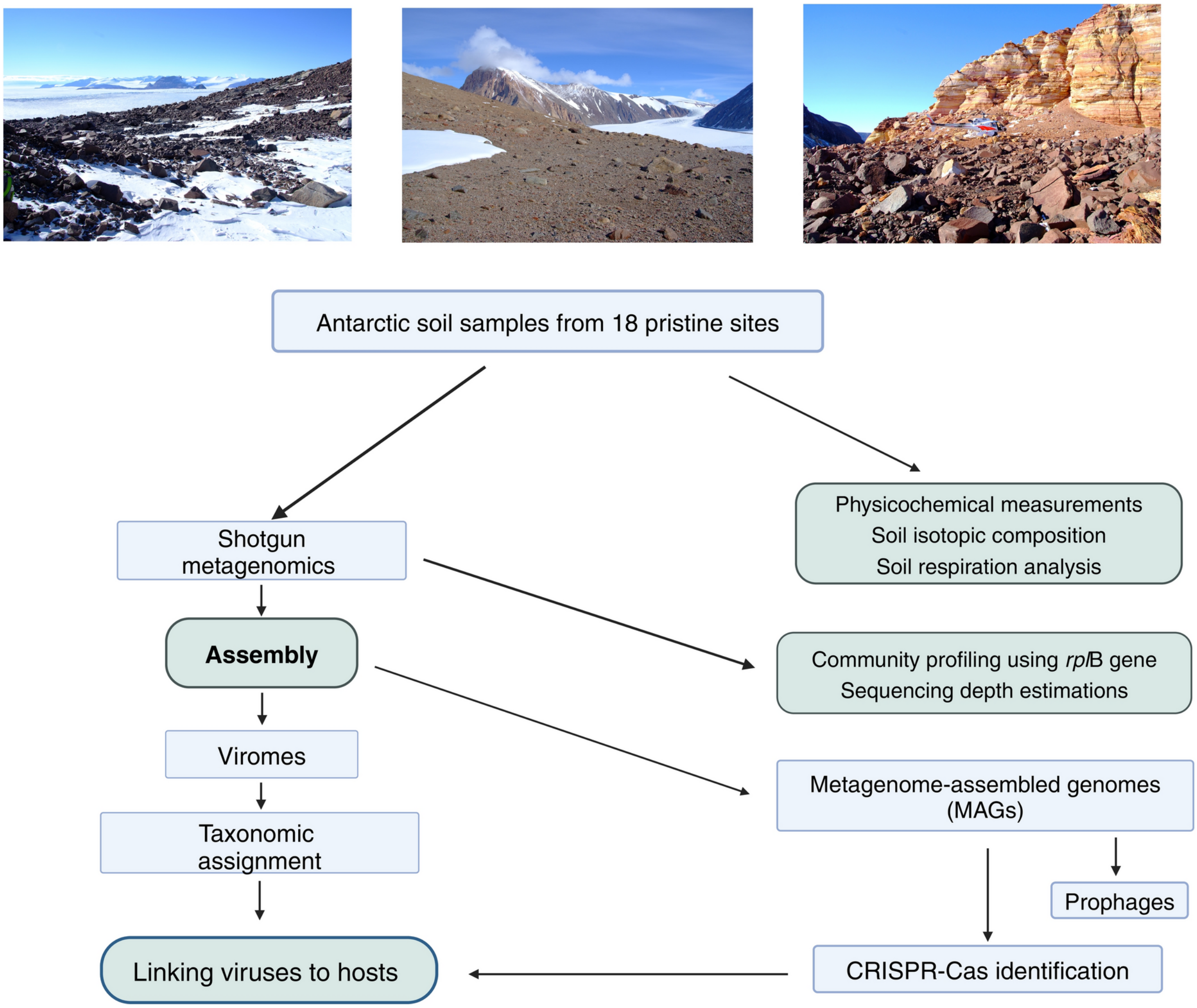
Antarctic soil samples from 18 pristine sites
Shotgun metagenomics
Physicochemical measurements
Soil isotopic composition
Soil respiration analysis
Assembly
Community profiling using *rpl*B gene
Sequencing depth estimations
Viromes
Metagenome-assembled genomes (MAGs)
Taxonomic assignment
Prophages
Linking viruses to hosts
CRISPR-Cas identification

Fig. 6. Sampling sites and methodology. Photographs showing three of the ice-free Antarctic sampling sites, from left to right; Mount Gran, Mackay Glacier site 3, and Pegtop Mountain (credit Prof. Don Cowan). Flow diagram indicates the broad methodology used in this study.

(ARGs) as hallmarks of resistance to bacterial antibiotic production[88]. The 18 retained MAGs were inspected for CRISPR-Cas repeats using MinCED[89] and for prophages (integrated viral genomes) using VirSorter v1.0.6[90]. A maximum likelihood phylogenomic tree, based on 49 core bacterial genes from our 18 MAGs and 100 reference genomes (RefSeq database), was built using FastTree2 v2.1.10. Phylogenomic trees were visualized in iTOL v6.34[91]. We used iRep to estimate bacterial replication rates[51] and gRodon2 to calculate growth rate[52].

Viral taxonomic analysis

We also explored each metagenomic assembly for bacteriophages using VirSorter v1.0.6[90]. Contigs were manually inspected for viral "hallmark" genes from categories 1 and 2 (complete viral contigs), and 4 and 5 (prophages). The quality of the predicted viral contigs was assessed using CheckV v1.0[92]. Contigs > 10 kb that were thought to be viral were then clustered using vConTACT2 v0.9.19[93] to establish a network of protein clusters among known phages from the Virome database. The edges of the network are significant gene-sharing similarities between contigs, which are represented as nodes. These uncultivated viral genomes (UViGs) were also inspected for possible auxiliary metabolic genes (AMGs) that could have been acquired from their host. We built phylogenetic trees using MAFFT v7.294b[94] of the AMGs identified in UViGs and their homologs in host MAGs to determine the possible origin of the gene. Finally we used the dbCAN2 web server[95] to identify glycosyl hydrolases (GH) and glycoside transferases (GT) in the MAGs. UViGs were clustered into viral OTUs (vOTUs) at 95% average nucleotide identity (ANI) and 85% alignment fraction (AF).

CRISPR spacer analysis and protein structure analysis using AlphaFold2

Metagenomes and MAGs were explored for the presence of adaptive immunity systems such as CRISPR-cas gene types using hmmsearch with (E-value $1e^{-05}$) against Cas gene profiles obtained from a study by Makarova, Wolf, Iranzo, Shmakov, Alkhnbashi, Brouns, Charpentier, Cheng, Haft and Horvath[33]. These were further assessed for the presence of innate immune response genes using RPS-BLAST (E-value $1e^{-05}$) against conserved domain databases (CDD) of clusters of orthologous groups (COGs) and protein families (Pfams)[96]. Results from these searches were manually filtered for the identification of CRISPR-Cas systems, toxin-antitoxins (TA), restriction-modification (RM), bacteriophage exclusion (BREX), abortive infection (Abi), defense island system associated with restriction-modification (DISARM), and other recently identified systems using a refined list of COG and Pfam identifiers reported to be associated with these defense mechanisms[33,97].

Cas reference sequences were extracted from UniProtKB/Swiss-Prot (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz; accessed 2021/05/25); a high-quality, manually annotated and non-redundant protein sequence database[98]. For each Cas protein family the reference amino acid sequences and our amino acid sequences were included in multiple sequence alignment using Mafft v7.294b with the "linsi" parameter specified[94]. Trees were constructed based on the multiple sequence alignments using IQ-TREE v2.1.2[99] with the "MFP" parameter invoked which uses ModelFinder[100] to determine the most appropriate model. Tree refinement was performed in iTOL v6.3[91]. For the Cas G + C plots we used the R v4.0.2 statistical environment while the G + C skew was calculated using iREP v1.10[51]. Cas protein structures were determined using AlphaFold2[101].

Bayesian analysis

Divergence time was estimated using Bayesian analyses. A multiple protein sequence alignment was constructed using GTDB-Tk v.1.7.0[87] and was based on the 120 GTDB core bacterial marker genes. The alignment included a set of 32 reference outgroups previously used to calibrate the crown bacteria[45,55]. BEAUti v.1.10.4[102] was used to specify parameters for Markov chain Monte Carlo (MCMC) tree analyses in BEAST v.1.10.4[103]. The multiple protein sequence alignment was imported and a Gamma Site Model with four categories was selected. The LG amino acid substitution model was used and a Relaxed Clock model with a Log Normal distribution. A Coalescent Constant Population tree prior was chosen and bacterial divergence (crown) was calibrated using a prior on the root of 3,453 million years ago (Ma) and a standard deviation of 60 Ma[45,55,104] with a Log Normal distribution specified. MCMC parameters were set at 10,000,000–40,000,000 Chain Lengths with a sampling frequency of 1,000. Tracer v.1.7.2[105] was used to assess convergence with burn in percentages between 10 and 80 to obtain the optimal effective sample size. A Maximum clade credibility tree and Mean heights were selected to produce a summarized target MCMC tree with TreeAnnotator v.1.10.4[103]. iTOL was used for final tree visualization and analysis[91].

Data availability

The quality-filtered, unassembled metagenomic sequences are available on the MG-RAST server under the accession numbers 4667018.3 to 4667036.3. All contigs longer than 200 bp from the assembled metagenomes are deposited on the NCBI under the BioProject PRJNA376086. Code for statistical analyses is available at https://github.com/SAmicrobiomes/.

Received: 29 August 2024; Accepted: 18 December 2024
Published online: 18 January 2025

References

1. Rodriguez-Valera, F. et al. Explaining microbial population genomics through phage predation. *Nat. Prec.*, 1–27 (2009).
2. Danovaro, R. et al. Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* **454**, 1084–1087 (2008).
3. Lara, E. et al. Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Sci. Adv.* **3**, e1602565 (2017).
4. Breitbart, M., Thompson, L. R., Suttle, C. A. & Sullivan, M. B. Exploring the vast diversity of marine viruses. *Oceanography* **20**, 135–139 (2007).
5. Nelson, A. R. et al. Wildfire-dependent changes in soil microbiome diversity and function. *Nat. Microbiol.* **7**, 1419–1430 (2022).
6. Bi, L. et al. Diversity and potential biogeochemical impacts of viruses in bulk and rhizosphere soils. *Environ. Microbiol.* **23**, 588–599 (2021).
7. Jansson, J. K. & Wu, R. Soil viral diversity, ecology and climate change. *Nat. Rev. Microbiol.* **21**, 1–16 (2022).
8. Liang, X. et al. Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. *Soil Biol. Biochem.* **144**, 107767 (2020).
9. Sokol, N. W. et al. Life and death in the soil microbiome: How ecological processes influence biogeochemistry. *Nat. Rev. Microbiol.* **7**, 1–16 (2022).
10. Jin, M. et al. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 1–15 (2019).
11. Anantharaman, K. et al. Sulfur oxidation genes in diverse deep-sea viruses. *Science* **344**, 757–760 (2014).
12. Lee, S., Sieradzki, E. T., Nicol, G. W. & Hazard, C. Propagation of viral genomes by replicating ammonia-oxidising archaea during soil nitrification. *ISME J.* **17**, 1–6 (2022).
13. Albright, M. B. et al. Experimental evidence for the impact of soil viruses on carbon cycling during surface plant litter decomposition. *ISME Commun.* **2**, 1–8 (2022).
14. Adriaenssens, E. M. et al. Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* **5**, 1–14 (2017).
15. Bezuidt, O. K. et al. Phages actively challenge niche communities in Antarctic soils. *Msystems* **5**, 1–12 (2020).
16. Van Goethem, M. W. et al. A reservoir of 'historical'antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome* **6**, 1–12 (2018).
17. Zablocki, O., Adriaenssens, E. M. & Cowan, D. Diversity and ecology of viruses in hyperarid desert soils. *Appl. Environ. Microbiol.* **82**, 770–777 (2016).
18. Kuzyakov, Y. & Mason-Jones, K. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol. Biochem.* **127**, 305–317 (2018).
19. García-Sastre, A. & Biron, C. A. Type 1 interferons and the virus-host relationship: A lesson in detente. *Science* **312**, 879–882 (2006).

20. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491 (2013).
21. Hampton, H. G., Watson, B. N. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
22. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
23. Hille, F. et al. The biology of CRISPR-Cas: Backward and forward. *Cell* **172**, 1239–1259 (2018).
24. Bikard, D. et al. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat. Biotechnol.* **32**, 1146–1150 (2014).
25. Barrangou, R. & Marraffini, L. A. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell* **54**, 234–244 (2014).
26. Koonin, E. V., Makarova, K. S., Wolf, Y. I. & Krupovic, M. Evolutionary entanglement of mobile genetic elements and host defence systems: Guns for hire. *Nat. Rev. Genet.* **21**, 119–131 (2020).
27. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117**, 119–128 (2015).
28. Faure, G. et al. CRISPR–Cas in mobile genetic elements: Counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
29. van Houte, S., Buckling, A. & Westra, E. R. Evolutionary ecology of prokaryotic immune mechanisms. *Microbiol. Mol. Biol. Rev.* **80**, 745–763 (2016).
30. Westra, E. R., Buckling, A. & Fineran, P. C. CRISPR–Cas systems: Beyond adaptive immunity. *Nat. Rev. Microbiol.* **12**, 317–326 (2014).
31. Mohanraju, P. et al. Alternative functions of CRISPR–Cas systems in the evolutionary arms race. *Nat. Rev. Microbiol.* **20**, 351–364 (2022).
32. Koonin, E. V. & Makarova, K. S. Mobile genetic elements and evolution of CRISPR-Cas systems: All the way there and back. *Genome Biol. Evol.* **9**, 2812–2825 (2017).
33. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: A burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
34. Pinilla-Redondo, R. et al. CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucl. Acids Res.* **50**, 4315–4328 (2022).
35. Tao, S. et al. Association of CRISPR-Cas System with the Antibiotic Resistance and Virulence Genes in Nosocomial Isolates of Enterococcus. *Infect. Drug Resist.*, 6939–6949 (2022).
36. Xu, C. et al. Programmable RNA editing with compact CRISPR–Cas13 systems from uncultivated microbes. *Nat. Methods* **18**, 499–506 (2021).
37. Collias, D. & Beisel, C. L. CRISPR technologies and the search for the PAM-free nuclease. *Nat. Commun.* **12**, 1–12 (2021).
38. Barrangou, R. & Horvath, P. A decade of discovery: CRISPR functions and applications. *Nat. Microbiol.* **2**, 1–9 (2017).
39. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary genomics of defense systems in archaea and bacteria. *Ann. Rev. Microbiol.* **71**, 233–261 (2017).
40. Klompe, S. E. & Sternberg, S. H. Harnessing "A Billion Years of Experimentation": The ongoing exploration and exploitation of CRISPR–Cas immune systems. *CRISPR J.* **1**, 141–158 (2018).
41. Makarova, K.S., Wolf, Y.I. & Koonin, E.V. Evolutionary Classification of CRISPR-Cas Systems. *CRISPR: Biol. Appl.*, 13–38 (2022).
42. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
43. Bernheim, A. & Sorek, R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Microbiol.* **18**, 113–119 (2020).
44. Hu, C. et al. Mechanism for Cas4-assisted directional spacer acquisition in CRISPR–Cas. *Nature* **598**, 515–520 (2021).
45. Ortiz, M. et al. Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils. *Proc. Natl. Acad. Sci.* **118**, e2025322118 (2021).
46. Collins, G. E. et al. Genetic diversity of soil invertebrates corroborates timing estimates for past collapses of the West Antarctic Ice Sheet. *Proc. Natl. Acad. Sci.* **117**, 22293–22302 (2020).
47. Coleine, C., Selbmann, L., Singh, B. K. & Delgado-Baquerizo, M. The poly-extreme tolerant black yeasts are prevalent under high ultraviolet light and climatic seasonality across soils of global biomes. *Environ. Microbiol.* **24**, 1988–1999 (2022).
48. Van Goethem, M.W. et al. Nutrient parsimony shapes diversity and functionality in hyper-oligotrophic Antarctic soils. *bioRxiv* (2020).
49. Xu, S., Wang, J., Guo, Z., He, Z. & Shi, S. Genomic convergence in the adaptation to extreme environments. *Plant Commun.* **1**, 1–14 (2020).
50. Kelley, J. L. et al. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat. Commun.* **5**, 4611 (2014).
51. Brown, C. T., Olm, M. R., Thomas, B. C. & Banfield, J. F. Measurement of bacterial replication rates in microbial communities. *Nat. Biotechnol.* **34**, 1256–1263 (2016).
52. Weissman, J. L., Hou, S. & Fuhrman, J. A. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc. Natil. Acad. Sci.* **118**, e2016810118 (2021).
53. Araujo, R. et al. Biogeography and emerging significance of Actinobacteria in Australia and Northern Antarctica soils. *Soil Biol. Biochem.* **146**, 107805 (2020).
54. Varliero, G. et al. Biogeographic survey of soil bacterial communities across Antarctica. *Microbiome* **12**, 1–22 (2024).
55. Albanese, D. et al. Pre-Cambrian roots of novel Antarctic cryptoendolithic bacterial lineages. *Microbiome* **9**, 1–15 (2021).
56. Braga, L. P. et al. Viruses direct carbon cycling in lake sediments under global change. *Proc. Natl. Acad. Sci.* **119**, e2202261119 (2022).
57. McKay, L. J. et al. Sulfur cycling and host-virus interactions in Aquificales-dominated biofilms from Yellowstone's hottest ecosystems. *ISME J.* **16**, 842–855 (2022).
58. Luo, X.-Q. et al. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 1–18 (2022).
59. Bradde, S., Nourmohammad, A., Goyal, S. & Balasubramanian, V. The size of the immune repertoire of bacteria. *Proc. Nat. Acad. Sci.* **117**, 5144–5151 (2020).
60. Jackson, S.A. et al. CRISPR-Cas: adapting to change. *Science* **356**, eaal5056 (2017).
61. Makarova, K. S., Anantharaman, V., Aravind, L. & Koonin, E. V. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biol. Direct* **7**, 1–10 (2012).
62. Garrett, S. C. Pruning and tending immune memories: spacer dynamics in the CRISPR array. *Front. Microbiol.* **12**, 664299 (2021).
63. Mangericao, T. C., Peng, Z. & Zhang, X. Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls. *BMC Syst. Biol.* **10**, 81–87 (2016).
64. Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinf.* **8**, 1–10 (2007).
65. Lopatina, A. et al. Metagenomic analysis of bacterial communities of Antarctic surface snow. *Front. Microbiol.* **7**, 1–13 (2016).
66. Yeo, C. C. GNAT toxins of bacterial toxin–antitoxin systems: acetylation of charged tRNAs to inhibit translation. *Mol. Microbiol.* **108**, 331–335 (2018).
67. Czub, M. P. et al. A Gcn5-related N-acetyltransferase (GNAT) capable of acetylating polymyxin B and colistin antibiotics in vitro. *Biochemistry* **57**, 7011–7020 (2018).

68. Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* **8**, e44248 (2019).
69. Makarova, K.S. & Koonin, E.V. Annotation and classification of CRISPR-Cas systems. *CRISPR*, 47–75 (2015).
70. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proc. Natl. Acad. Sci.* **114**, E7358–E7366 (2017).
71. Phale, P. S., Shah, B. A. & Malhotra, H. Variability in assembly of degradation operons for naphthalene and its derivative, carbaryl, suggests mobilization through horizontal gene transfer. *Genes* **10**, 569 (2019).
72. Daubin, V., Lerat, E. & Perrière, G. The source of laterally transferred genes in bacterial genomes. *Genome Biol.* **4**, 1–12 (2003).
73. Harrington, L. B. et al. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839–842 (2018).
74. Yan, W. X. et al. Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88–91 (2019).
75. Koonin, E. V. & Makarova, K. S. Mobile Genetic Elements and Evolution of CRISPR-Cas Systems: All the Way There and Back. *Genome Biol Evol* **9**, 2812–2825 (2017).
76. Klages, J. P. et al. Temperate rainforests near the South Pole during peak Cretaceous warmth. *Nature* **580**, 81–86 (2020).
77. Altae-Tran, H. et al. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382**, eadi1910 (2023).
78. Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R. & Konstantinidis, K.T. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *MSystems* **3** (2018).
79. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
80. Bushnell, B. BBTools software package. URL http://sourceforge.net/projects/bbmap **578**, 579 (2014).
81. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genom. Sci.* **10**, 1–4 (2015).
82. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
83. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 1–18 (2015).
84. Kang, D. D. et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
85. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
86. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
87. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
88. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research* **48**, 517–525 (2020).
89. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinf.* **8**, 1–8 (2007).
90. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
91. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucl. Acids Res.* **49**, 293–296 (2021).
92. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature biotechnology* **39**, 578–585 (2021).
93. Jang, H. B. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology* **37**, 632–639 (2019).
94. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780 (2013).
95. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucl. Acids Res.* **46**, 95–101 (2018).
96. Bateman, A. et al. The Pfam protein families database. *Nucl. Acids Res.* **32**, 138–141 (2004).
97. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
98. UniProt: the universal protein knowledgebase in 2021. *Nucl. Acids Res.* **49**, D480-D489 (2021).
99. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
100. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
101. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
102. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
103. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).
104. Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
105. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

Acknowledgements

We wish to thank the Centre for High Performance Computing and the University of Pretoria's Centre for Bioinformatics and Computational Biology for server access. TRN was supported by the Office of Science Early Career Research Program Office of Biological and Environmental Research, Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory.

Author contributions

T.P.M designed the research with input from M.W.V.G, R.P., and O.K.I.B. M.W.V.G., R.P., O.K.I.B., S.V. analyzed the data. D.A.C, I.D.H., D.W.H., T.A., G.H., S.W., T.R.N., W.K., D.D., Y.V.d.P., M.D.B. and T.P.M. coordinated field and laboratory operations. The manuscript was written by M.W.V.G. and T.P.M. with contributions from all co-authors.

Funding
We gratefully acknowledge the National Research Foundation (NRF; Grant ID 118981), the South African National Antarctic Programme (SANAP 110717), and the University of Pretoria for funding. We acknowledge Antarctica NZ for logistics and support and thank Professor Ian Hogg (University of Waikato, NZ) as the leader of Event K024.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-83942-y.

Correspondence and requests for materials should be addressed to T.P.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.