

BRIEF REPORT

Open Access



Highly diverse and unknown viruses may enhance Antarctic endoliths' adaptability

Cassandra L. Ettinger^{1*}, Morgan Saunders^{2,3}, Laura Selbmann⁴, Manuel Delgado-Baquerizo^{5,6}, Claudio Donati⁷, Davide Albanese⁷, Simon Roux⁸, Susannah Tringe⁸, Christa Pennacchio⁸, Tijana G. del Rio⁸, Jason E. Stajich^{1,9} and Claudia Coleine^{1,4,5*}

Abstract

Background Rock-dwelling microorganisms are key players in ecosystem functioning of Antarctic ice free-areas. Yet, little is known about their diversity and ecology, and further still, viruses in these communities have been largely unexplored despite important roles related to host metabolism and nutrient cycling. To begin to address this, we present a large-scale viral catalog from Antarctic rock microbial communities.

Results We performed metagenomic analyses on rocks from across Antarctica representing a broad range of environmental and spatial conditions, and which resulted in a predicted viral catalog comprising > 75,000 viral operational taxonomic units (vOTUS). We found largely undescribed, highly diverse and spatially structured virus communities which had predicted auxiliary metabolic genes (AMGs) with functions indicating that they may be potentially influencing bacterial adaptation and biogeochemistry.

Conclusion This catalog lays the foundation for expanding knowledge of virosphere diversity, function, spatial ecology, and dynamics in extreme environments. This work serves as a step towards exploring adaptability of microbial communities in the face of a changing climate.

*Correspondence:
Cassandra L. Ettinger
cassande@ucr.edu
Claudia Coleine
coleine@unitus.it

¹ Department of Microbiology and Plant Pathology, University of California, Riverside, CA, USA

² Department of Biological Sciences, California State Polytechnic University, Pomona, CA, USA

³ The University of Arizona, Tucson, AZ, USA

⁴ Department of Ecological and Biological Sciences, University of Tuscia, Viterbo, Italy

⁵ Laboratorio de Biodiversidad Y Funcionamiento Ecosistémico, Instituto de Recursos Naturales Y Agrobiología de Sevilla (IRNAS), CSIC, Av. Reina Mercedes 10, Seville E-41012, Spain

⁶ Unidad Asociada CSIC-UPO (BioFun), Universidad Pablo de Olavide, Seville 41013, Spain

⁷ Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, San Michele All'Adige 38098, Italy

⁸ Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720, USA

⁹ Institute for Integrative Genome Biology, University of California, Riverside, CA, USA



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Viruses are among the most prevalent entities on our planet, with the ability to infect organisms across all domains [1]. Sequencing advances are reshaping understanding of viral diversity across Earth's diverse ecosystems, leading to a remarkable expansion of viral catalogs [1–6]. It is becoming clear that viruses play key roles in global biogeochemical cycles through the modulation of host population dynamics, and that the better-studied pathogenic viruses represent only a small fraction of the virosphere [7–9]. Further, through auxiliary metabolic genes (AMGs), some viruses can directly impact host metabolism to improve fitness [10], including in terrestrial ecosystems characterized by extreme conditions (e.g., oligotrophy, aridity, high, or low temperature).

Antarctic ice-free areas include several of the most inhospitable regions on Earth, among which is the Mars counterpart [11]: the McMurdo Dry Valleys. In these locations, where rocks represent the main substratum, active life is possible for only a few specialized microorganisms; they survive by dwelling in porous rocks, forming self-sustaining ecosystems called endolithic communities [12, 13]. These microorganisms are the primary life-forms present assuring the balance and functionality of these otherwise inert ecosystems. Recent studies have shed light on their biodiversity and adaptation, particularly the evolution of new and peculiar taxa spanning bacteria, fungi, and archaea [13–16]. However, the ecology and distribution of viral diversity from these communities remain wholly unknown and, to date, viral studies have instead focused on Antarctic freshwater lakes [17–19], surrounding oceans [20–22], and soils [23–26].

Here, we provide a large-scale viral catalog from 191 Antarctic endolith metagenomes. We sampled 37 localities across a broad range of environmental (e.g., 4 rock typologies, different altitudes and sun exposure) and spatial conditions (i.e., Antarctic Peninsula, Northern Victoria Land, and McMurdo Dry Valleys) (Table S1; Fig. 2A). We aimed to (i) untangle viral diversity in these communities, (ii) predict AMGs and how they may drive the fitness of their hosts, and (iii) explore ecological patterns (e.g., biogeography). This catalog is the first step toward understanding the role of viruses in the coldest and driest region on Earth. This information is also critical for elucidating the possible role of viruses in whole community adaptation in a dry ecosystem that will expand owing to global change [27].

Methods

Study area

One hundred ninety-one rocks colonized by endolithic communities were collected in thirty-eight sites in Antarctica including Antarctic Peninsula ($n=3$),

McMurdo Dry Valleys, Southern Victoria Land ($n=80$), and Northern Victoria Land ($n=108$) during more than 20 years of Italian Antarctic Expeditions. Different rock typologies (sandstone $n=141$, granite $n=43$, quartz $n=5$, and basalt/dolerite $n=2$) were sampled. Samples were collected along a latitudinal transect ranging from $-62.10008-58.51664$ to $-77.874-160.739$ at different environmental conditions namely sun exposure (northern sun exposed and southern shady rocks) and an altitudinal transect from sea level to 3100 m above sea level (a.s.l.) to provide a comprehensive overview of Antarctic endolithic diversity (Table S1). The presence of endolithic colonization was assessed by direct observation in situ. Rocks were excised using a geologic hammer and sterile chisel, and rock samples were preserved in sterile plastic bags, transported, and stored at $-20\text{ }^{\circ}\text{C}$ in the Culture Collection of Antarctic fungi of the Mycological Section of the Italian Antarctic National Museum (MNA-CCFEE), until downstream analysis.

Study data

In total, the dataset included 191 metagenomes, of which 100 have been assembled as described in Albanese et al. [14]. The remaining metagenomes were generated, sequenced, and assembled as described below. The final metagenomic set represented 149,585,625 metagenomic contigs.

DNA extraction, library preparation, and sequencing

Total community DNA was extracted from 1 g of crushed rocks using DNeasy PowerSoil Pro Kit (Qiagen, Germany), quality checked by electrophoresis using a 1.5% agarose gel and Nanodrop spectrophotometer (ThermoFisher, USA) and quantified using the Qubit dsDNA HS Assay Kit (Life Technologies, USA) according to Coleine et al. [13]. Shotgun metagenomic sequencing paired-end libraries were constructed by using Next Ultra DNA library prep kits and sequenced as 2×150 bp using the Illumina NovaSeq platform (Illumina Inc., San Diego, CA, USA) at the Edmund Mach Foundation (San Michele all'Adige, Italy) and at the DOE Joint Genome Institute (JGI).

Sequencing reads preparation and assembly

The metashot/mag-illumina v2.0.0 workflow (<https://github.com/metashot/mag-illumina>, parameters: `–metaspades_k 21,33,55,77,99`) was used to perform raw reads quality trimming and filtering, and assembly on the metagenomic samples. In brief, adapter trimming, contaminant (artifacts and spike-ins) and quality filtering were performed using BBDuk (BBMap/BBTools v38.79, <https://sourceforge.net/projects/bbmap/>). During the quality filtering procedure i) raw reads were quality-trimmed to Q6 using the Phred

algorithm; ii) reads that contained 4 or more “N” bases, had an average quality below 10, shorter than 50 bp or under 50% of the original length were removed. Samples were then assembled individually with SPAdes v3.15.1 [28] (parameters `-meta -k 21,33,55,77,99`).

Identification and clustering of viral genomes

Using a workflow similar to Guo et al. [29], viral sequences were identified separately for each of the 191 metagenomic assemblies using VirSorter2 v. 2.2.3 [30] using `-min-length 5000, -min-score 0.5, and -include-groups dsDNA phage,NCLDV,RNA,ssDNA,lavidaviridae`. CheckV v0.8.1 [31] was run on the VirSorter2 predicted viral sequences using the “end_to_end” workflow. VirSorter2 was then run again on the viral sequences from CheckV workflow with the `-prep-for-dramv` option. DRAM-v v. 1.2.2 [32] was then used to “annotate” sequences against databases downloaded on August 3rd, 2021 including Viral RefSeq, KOfam [33], Pfam [34], dbCAN v.9 [35], MEROPS [36], and VOGDB (<http://vogdb.org/>). DRAM-v was then used to “distill” annotations into predicted AMGs for phage.

Viral sequences from all assemblies were combined and clustered into 95% similarity viral operational taxonomic units (vOTUs) using CD-HIT v. 4.8.1 [37] with the following parameters: `-c 0.95 -aS 0.85 -M 0 -d 0`. Prodigal v. 2.6.3 [38] was used to predict open reading frames in vOTUs using the `-p meta` option. VContact2 v. 0.9.19 was then run on predicted proteins from phage vOTUs and predicted proteins from the INPHARED August 2022 viral reference database to generate viral clusters (VCs) based on gene-sharing networks [39, 40]. We assigned taxonomy to phage vOTUs based on VC membership as in Santos-Medellin et al. [41]. Predicted viral sequences and 95% similarity vOTUS are archived on Zenodo [42].

Viral host-prediction

Hosts were predicted for the phage sequences identified using (i) a database of complete genomes from NCBI RefSeq, and (ii) a previously published database of representative metagenome-assembled genomes (MAGs) from Antarctic endolith samples. To produce (i), we used “ncbi-genome-download” to download all complete bacterial ($n=25,984$) and archaeal ($n=416$) genomes, as of April 7, 2022, from NCBI RefSeq [43]. For (ii), we downloaded MAGs from Zenodo (<https://doi.org/10.5281/zenodo.7313591>). We then used NCBI BLAST 2.12.0+ to convert these two databases into blast databases using “makeblastdb” and used “blastn” to compare vOTUs to these databases [44]. We filtered the blastn results in R based on existing thresholds [45–47]. Briefly, database matches had to share ≥ 2000 bp region with $\geq 70\%$ sequence identity to the viral sequence and needed to have a bit score of ≥ 50 and minimum e value of 0.001.

Further to ensure matches did not represent partial or entirely viral contigs when searching against the MAG database, matches had to cover $< 50\%$ of the total MAG sequence length. As in Korthari et al. [46], only the top 5 hits matching these thresholds were considered, with host predictions made at each taxonomic level only if the taxonomy of all hits were in agreement. Discrepancies resulted in no host prediction for that taxonomic level. We then combined host predictions from both the RefSeq and MAG databases together; if there were discrepancies between the two databases, we defaulted to the MAG-based prediction.

Ecological analysis of vOTUs

We mapped reads from each metagenome to vOTUs using BMap with a minimum sequence similarity of 90% to quantify vOTU relative abundance [48]. We then used SAMtools to convert resulting sam files to bam files and genomecov from BEDTools to obtain coverage information for each vOTU across each metagenome [49, 50]. We then used bamM to parse bam files and calculate the trimmed pileup coverage (tpmean), which we used here in our analysis of viral relative abundance [51]. We removed vOTUs which displayed $< 75\%$ coverage over the length of the viral sequence and viral sequences < 10 kbp in length prior to downstream analyses in R [52]. Thresholds for analysis of vOTUs were based on community guidelines for length (i.e., ≥ 10 kbp), similarity (i.e., $\geq 95\%$ similarity), and detection (i.e., $\geq 75\%$ of the viral genome length covered $\geq 1 \times$ by reads at $\geq 90\%$ average nucleotide identity) [53, 54]. To be conservative, we also removed vOTUs with a CheckV quality score of “not-determined” prior to downstream analysis. The viral abundance (tpmean), quality, taxonomy, and annotation results were imported, analyzed, and visualized in R using many packages including tidyverse and phyloseq [55, 56]. Analysis scripts associated with this study are on GitHub and archived in Zenodo [57].

To compare viral diversity between metagenomes (i.e., beta diversity), we calculated the Hellinger distance, the Euclidean distance of Hellinger transformed abundance data. We performed Hellinger transformations using the transform function in the microbiome R package, calculated the Hellinger distance using the ordinate function in phyloseq, and then visualized these distances using principal-coordinate analysis (PCoA). We performed permutational multivariate analyses of variance (PERMANOVAs) with 9,999 permutations to test for significant differences in mean centroids using the model: Distance \sim Site + Rock type. Models were tested with “by=margins” and “by=terms” with all sequential combinations. We ran the ordistep and ordiR2step functions to help assess optimal parameters to include in the model. Since PERMANOVA tests are sensitive to differences in group dispersion, we

also tested for significant differences in mean dispersions using the betadisper and permutest functions from the vegan package in R with 9,999 permutations.

To test for correlations between viral community distances (Hellinger distances) and geographic distances, we first subset the data to exclude metagenomes from the Antarctic Peninsula, and to account for variation between rock types, subset the data to include only metagenomes representing sandstone samples. We calculated geographical distances between metagenomes using the distm function in the geosphere package in R. We performed Mantel tests in the vegan R package to assess correlations between the community and geographic distances using a Spearman correlation and 9999 permutations. Mantel tests were repeated with exclusion of community distances when the geographic distance was zero to assess if patterns persisted in the absence of data from the same site.

Results and discussion

Using VirSorter2 [30], we predicted 101,085 viral sequences. We clustered these at 95% average nucleotide identity into 76,984 vOTUS [37]; we further used

VContact2 [39] with INPHARED [40] reference genomes to cluster phage vOTUs into 7598 VCs, which approximate genus-level groupings based on gene-sharing networks. To keep analysis focused on the most robust catalog, we filtered this collection using community thresholds for length, detection, and quality (see “Methods” section) [31, 53, 54]. The final viral catalog represented 14,796 viral sequences (Table S2; 76 complete, 341 high-quality, 1539 medium-quality, 12,840 low-quality), including 2,695 prophage, which clustered into 11,806 vOTUs, of which 5743 phage vOTUs (7309 sequences) were successfully placed in 2286 VCs; the final catalog was predicted to predominantly be dsDNA phage, though 15.2% of vOTUs may represent eukaryotic viruses (i.e. nucleocytoplasmic large DNA viruses).

Our findings may indicate that Antarctic rock communities host highly diverse and novel phage populations, with only 1.8% (41 out of 2,286) of the VCs including reference sequences. The remaining 98.2% were unique VCs (i.e., did not include reference genomes), and could represent novel phage genera, greatly expanding the known diversity of viruses. Of

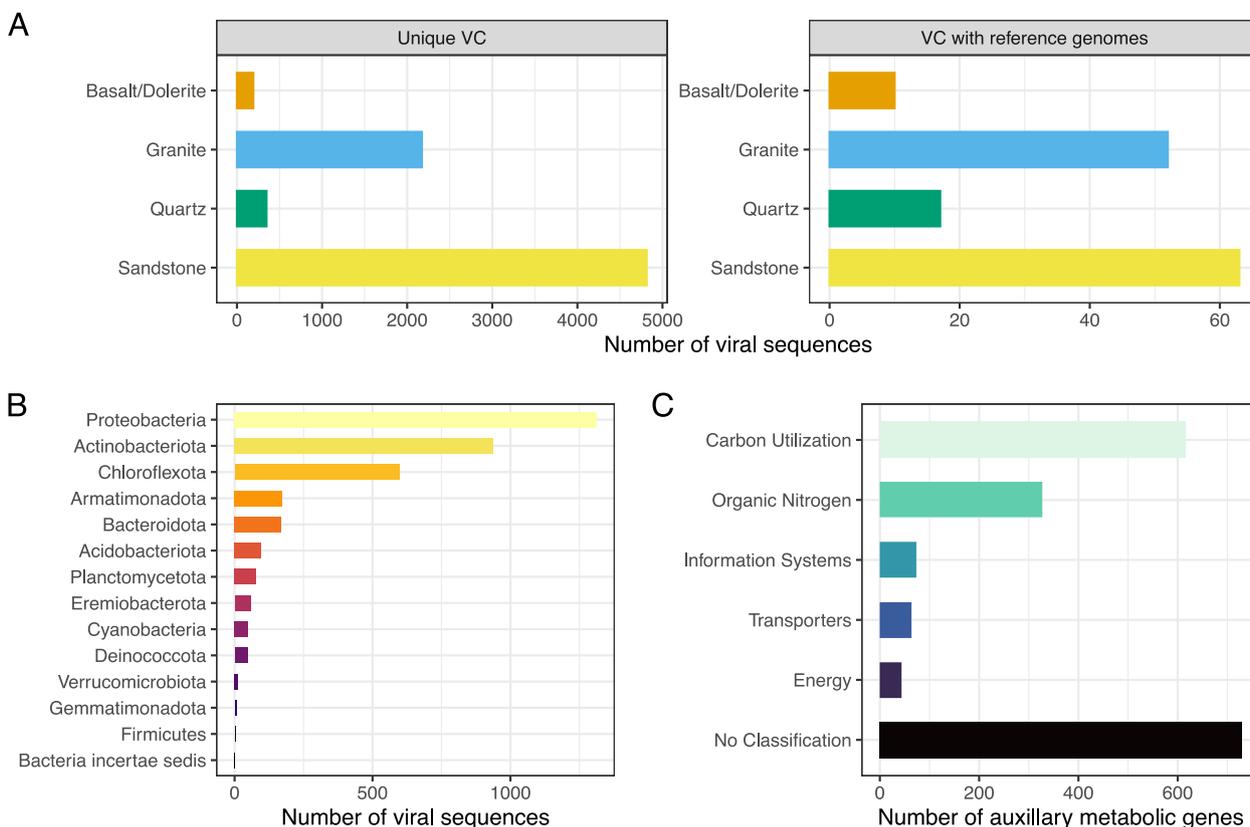


Fig. 1 Antarctica is an underappreciated source of phage novelty. **A** Bar charts displaying the number of viral sequences placed in VCs colored by rock type (sandstone $n = 141$, granite $n = 43$, quartz $n = 5$, and basalt/dolerite $n = 2$) and divided by whether the VC is clustered with reference genomes. **B** Bar chart displaying host predictions colored by predicted host phylum. **C** Bar chart showing the number of predicted phage AMGs summarized by DRAM-v distilled metabolic categories

the 41 VCs that did include reference genomes, the majority were assigned to the *Caudoviricetes* class (formerly *Caudovirales* order) of tailed double-stranded DNA bacteriophage (Figure S1). Many genomes have not yet been reclassified, leaving viral taxonomy in flux; under the new schema, most of the 41 VCs are unclassified [58]. The majority of unique VCs are represented by viral sequences from sandstone communities (Fig. 1A), which represents an optimum substratum, in terms of rock traits (e.g., porosity), for endolithic colonization [59], but is also the most represented substratum in this work.

We further established host–virus linkages using NCBI BLAST against complete bacteria and archaea

genomes from RefSeq, and Antarctic endolithic bacterial and archaeal metagenome-assembled genomes (MAGs) (see “Methods” section) [44–46] to explore the potential effects of viruses on host fitness, such as host-cell reprogramming through AMGs [60]. While we were unable to predict hosts for the majority of viral sequences (only 23.94% had a host prediction), we observed that Proteobacteria, Actinobacteriota, and Chloroflexota were the most commonly predicted host phyla (Fig. 1B), which are thought to be core members of these communities [14, 61, 62]. Using predictions against the Antarctic MAGs, we predicted hosts for an additional 16.5% of viral sequences compared to the 7.48% predicted using RefSeq alone (Figure S2).

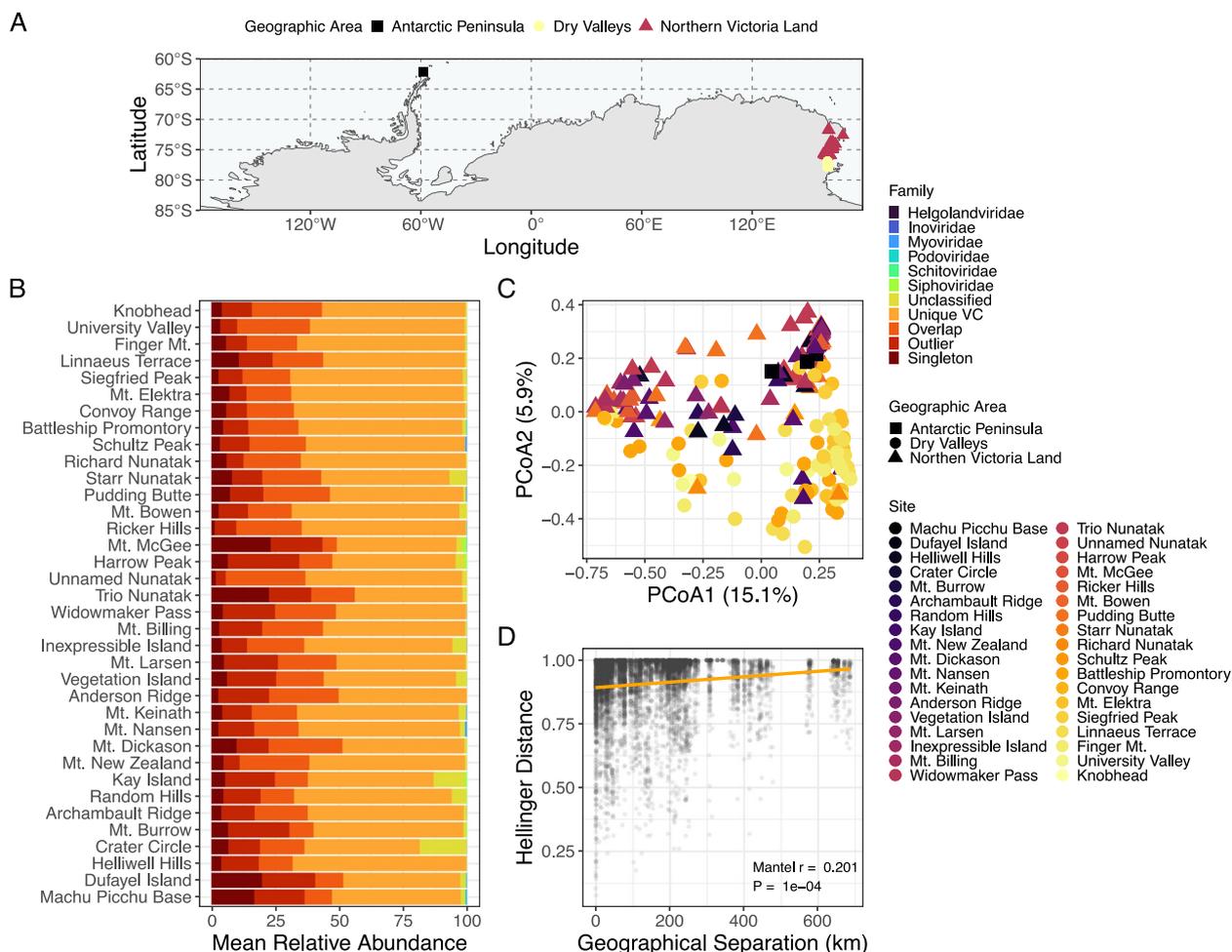


Fig. 2 Spatial structuring of viral communities in Antarctic rocks. **A** Map showing collection sites with shapes and colors representative of the broad geographic area. **B** Stacked bar charts displaying the mean relative abundance of phage vOTUs at each site colored by predicted viral families. Sequences that were clustered into VCs with reference data are labeled by their taxonomy, sequences clustered without reference genomes are labeled “Unique VC”, while the rest are labeled based on their VContact2 status (i.e., singleton [share few or no genes with other genomes], overlap [share genes with genomes in multiple VCs], or outlier [share genes, but cannot confidently be placed in a VC]). **C** Principal-coordinate analysis (PCoA) visualization of Hellinger distances of viral communities. Samples are colored by site, with sites ordered by latitude, and have shapes based on geographic areas. **D** A scatter plot depicting a significant relationship between sandstone viral community beta diversity (Hellinger distance) and geographical distance (km) between sites

We then sought to improve understanding on the functional profiles of retrieved phages using DRAM-v (Table S3) [32]. Notably, this catalog, which comprises metabolic novelty (39.3% of DRAM-v predicted AMG had no distilled classification), may complement other available resources, which have largely been limited to coverage of human-related microbiomes (e.g. Li et al. [63]). Within identified functions, we found putative phage AMGs related to carbon, energy, and nitrogen metabolisms (Fig. 1C). Specifically, within carbohydrate metabolism, glycoside hydrolases, glycosyltransferases, and carbohydrate-binding domains predominated. Within nitrogen metabolism, methionine degradation was the most prevalent module, and within energy, the dominant modules were related to electron transport and photosynthesis. This highlights the need to connect vOTUs to Antarctic MAGs [14] and to implement complementary techniques (e.g., single-cell genomics) to provide a deeper understanding of virus-bacteria dynamics. More importantly, these findings suggest a possible complex role for viruses in element biogeochemical cycles in the rocks of Antarctica, which have traditionally been considered devoid of life.

Given the geographic spread of sampling (see “Methods” section and Table S1; Fig. 2A), we assessed whether this catalog could be useful to answer ecological questions related to viral community dynamics. While the dominant vOTUs at each site were taxonomically unclassified and largely members of unique VCs and thus possible novel genera (Fig. 2B), when investigating between-sample diversity (beta diversity) we observed a significant pattern related to site specificity (Fig. 2C; PERMANOVA, $p < 0.001$). Further, we detected a significant correlation with geographic distance in sandstone communities (Fig. 2D; Mantel $r = 0.197$, $p < 0.001$), such that communities are more dissimilar with increasing distance. Combined these results indicate possible latitudinal spatial structuring of viral communities. In further support of this, we were able to detect only 41.0% of vOTUs at more than one site, with 29.4% of vOTUs detected across two or more geographic regions and only 1.45% detected across all regions. Of the vOTUs detected across all regions, the majority were in unique VCs (66.7%) and none were in VCs with reference data. We hypothesize that this viral spatial structuring reflects the reported dispersion limitation and local composition and adaptation of hosts in these communities [14, 15]. Similar spatial structuring has also been observed in grassland soil viromes, purportedly as a result of local assembly dynamics [41, 64].

Conclusions

This study represents the most exhaustive geographic endeavor to date to capture the viral genomic diversity across ice-free regions of Antarctica and the first

large-scale effort to explore the virosphere in endolithic communities. This catalog is a comprehensive repository for exploring the diversity, function, spatial ecology, and host-virus dynamics of this enigmatic continent. We also unveiled a potential influence of some viruses on carbon, energy, and nitrogen metabolism under conditions of oligotrophy up to the limit for life sustainability. Finally, this work may serve in the future as an important first step towards exploring adaptability of microbial communities in extreme conditions on Earth.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-023-01554-6>.

Additional file 1: Table S1. Sample information and viral identification statistics from Antarctic metagenomes. **Table S2.** Final viral catalog sequence information. **Table S3.** DRAM-v distilled annotations for predicted phage AMGs. **Figure S1.** Taxonomic classification of viral clusters (VCs) that include reference genomes. **Figure S2.** Comparison of host predictions between MAG and RefSeq databases.

Acknowledgements

Not applicable.

Authors' contributions

C.L.E., J.E.S., and C.C. conceived and designed the study. L.S. collected the samples. C.P., T.G.R. and S.T. oversaw and managed the metagenome sequencing and standard analysis. C.L.E. performed bioinformatic and statistical analysis with contributions from J.E.S. and M.S.. C.L.E. and C.C. interpreted results with contributions from M.S. and S.R.. C.L.E. and C.C. wrote the paper with contributions from all co-authors. The author(s) read and approved the final manuscript.

Funding

C.L.E. is supported by the National Science Foundation (NSF) under a NSF Ocean Sciences Postdoctoral Fellowship (Award No. 2205744). M.S. was supported by the NSF REU The National Summer Undergraduate Research Project, Award No. 2149582. C.C. is supported by the European Commission under the Marie Skłodowska-Curie Grant Agreement No. 702057 (DRYLIFE). C.C. and L.S. wish to thank the Italian National Antarctic Research Program for funding sampling campaigns and research activities in Italy in the frame of PNRA projects. The Italian Antarctic National Museum (MNA) is kindly acknowledged for financial support to the Mycological Section of the MNA and for providing rock samples used in this study stored in the Culture Collection of Antarctic fungi (MNA-CCFEE), University of Tuscia, Italy. M.D.B. is supported by a project from the Spanish Ministry of Science and Innovation (PID2020-115813RA-I00), and a project of the Fondo Europeo de Desarrollo Regional (FEDER) and the Consejería de Transformación Económica, Industria, Conocimiento y Universidades of the Junta de Andalucía (FEDER Andalucía 2014–2020 Objetivo temático '01–Refuerzo de la investigación, el desarrollo tecnológico y la innovación') associated with the research project P20_00879 (ANDABIOMA). J.E.S. is a CIFAR fellow in the Fungal Kingdom: Threats and Opportunities program. Data analyses performed at the High-Performance Computing Cluster at the University of California Riverside in the Institute of Integrative Genome Biology were supported by NSF grant DBI-1429826 and NIH grant S10-OD016290. Part of this work (proposals <https://doi.org/10.46936/10.25585/60000791> and <https://doi.org/10.46936/fics.proj.2020.51548/60000213>) was conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Availability of data and materials

Metagenomic raw data are available under the NCBI accession numbers listed in Supplementary Table S1. Analysis scripts and intermediate data files associated with this study are on GitHub (https://github.com/stajichlab/Antarctic_Virus_Discovery) and archived in Zenodo (<https://doi.org/10.5281/zenodo.7374327>). Fasta files representing the entire catalog of predicted viral sequences and 95% similarity vOTUS are archived and available on Zenodo (<https://doi.org/10.5281/zenodo.7245811>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 12 January 2023 Accepted: 19 April 2023

Published online: 08 May 2023

References

- Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. *Nature*. 2016;536:425–30.
- Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol*. 2018;3(8):870–80.
- Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, et al. Marine DNA viral macro-and microdiversity from pole to pole. *Cell*. 2019;177(5):1109–23.
- Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. *Cell Host Microbe*. 2020;28(5):724–40.
- Roux S, Paez-Espino D, Chen IM, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res*. 2021;49(D1):D764–75.
- Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell*. 2022;185(21):4023–37.
- Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol*. 2010;18:11–9.
- Jin M, Guo X, Zhang R, Qu W, Gao B, Zeng R. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome*. 2019;7:58.
- Hurwitz BL, Westveld AH, Brum JR, Sullivan MB. Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc Natl Acad Sci U S A*. 2014;111:10714–9.
- Breitbart M, Thompson L, Suttle C, Sullivan M. Exploring the vast diversity of marine viruses. *Oceanography*. 2007;20:135–9.
- Cockell CS, Bush T, Bryce C, Direito S, Fox-Powell M, Harrison JP, et al. Habitability: a review. *Astrobiology*. 2016;16:89–117.
- Friedmann EI. Endolithic microorganisms in the antarctic cold desert. *Science*. 1982;215:1045–53.
- Coleine C, Biagioli F, de Vera JP, Onofri S, Selbmann L. Endolithic microbial composition in Helliwell Hills, a newly investigated Mars-like area in Antarctica. *Environ Microbiol*. 2021;23:4002–16.
- Albanese D, Coleine C, Rota-Stabelli O, Onofri S, Tringe SG, Stajich JE, et al. Pre-Cambrian roots of novel Antarctic cryptoendolithic bacterial lineages. *Microbiome*. 2021;9:63.
- Archer SDJ, de los Ríos A, Lee KC, Niederberger TS, Craig Cary S, Coyne KJ, et al. Endolithic microbial diversity in sandstone and granite from the McMurdo Dry Valleys, Antarctica. *Polar Biol*. 2017;40:997–1006.
- de la Torre JR, Goebel BM, Friedmann EI, Pace NR. Microbial diversity of cryptoendolithic communities from the McMurdo Dry Valleys, Antarctica. *Appl Environ Microbiol*. 2003;69:3858–67.
- López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcami A. High diversity of the viral community from an Antarctic lake. *Science*. 2009;326:858–61.
- Prado T, Brandão ML, Fumian TM, Freitas L, Chame M, Leomil L, et al. Virome analysis in lakes of the South Shetland Islands, Antarctica - 2020. *Sci Total Environ*. 2022;852:158537.
- Zawar-Reza P, Argüello-Astorga GR, Kraberger S, Julian L, Stainton D, Broady PA, et al. Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect Genet Evol*. 2014;26:132–8.
- Alarcón-Schumacher T, Guajardo-Leiva S, Antón J, Díez B. Elucidating viral communities during a phytoplankton bloom on the West Antarctic Peninsula. *Front Microbiol*. 2019;10:1014.
- Miranda JA, Culley AI, Schvarcz CR, Steward GF. RNA viruses as major contributors to Antarctic viroplankton. *Environ Microbiol*. 2016;18:3714–27.
- Gong Z, Liang Y, Wang M, Jiang Y, Yang Q, Xia J, et al. Viral diversity and its relationship with environmental factors at the surface and deep sea of Prydz Bay, Antarctica. *Front Microbiol*. 2018;9:2981.
- Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, et al. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. *Appl Environ Microbiol*. 2014;80:6888–97.
- Adriaenssens EM, Kramer R, Van Goethem MW, Makhalyane TP, Hogg I, Cowan DA. Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome*. 2017;5:83.
- Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, et al. Niche-dependent genetic diversity in Antarctic metaviromes. *Bacteriophage*. 2014;4:e980125.
- Bezuidt OKI, Lebre PH, Pierneef R, León-Sobrino C, Adriaenssens EM, Cowan DA, et al. Phages actively challenge niche communities in antarctic soils. *mSystems*. 2020;5:e00234.
- Jansson JK, Wu R. Soil viral diversity, ecology and climate change. *Nat Rev Microbiol*. 2023;21:296–311. <https://doi.org/10.1038/s41579-022-00811-z>.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–77.
- Guo J, Vik D, Pratama AA, Roux S, Sullivan M. Viral sequence identification SOP with VirSorter2. *protocols.io*. 2021. <https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-btv8nn9w>.
- Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. 2021;9:37.
- Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol*. 2021;39:578–85.
- Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res*. 2020;48:8883–900.
- Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*. 2020;36:2251–2.
- El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. *Nucleic Acids Res*. 2019;47:D427–32.
- Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46:W95–101.
- Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. *Nucleic Acids Res*. 2010;38 Database issue:D227–33.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–9.
- Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119.
- Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol*. 2019;37:632–9.
- Cook R, Brown N, Redgwelt T, Rihtman B, Barnes M, Clokie M, et al. INFrA-structure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. *PHAGE*. 2021;2:214–23.

41. Santos-Medellin C, Zinke LA, Ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *ISME J*. 2021;15:1956–70.
42. Ettinger C, Stajich J, Coleine C. Antarctic rock viral catalog. 2022. <https://zenodo.org/record/7245811>.
43. kblin/ncbi-genome-download: Scripts to download genomes from the NCBI FTP servers. <https://github.com/kblin/ncbi-genome-download>.
44. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:1–9.
45. Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol Rev*. 2016;40:258–72.
46. Kothari A, Roux S, Zhang H, Prieto A, Soneja D, Chandonia J-M, et al. Ecogenomics of groundwater phages suggests niche differentiation linked to specific environmental tolerance. *mSystems*. 2021;6:e0053721.
47. Nayfach S, Roux S, Seshadri R, Udwy D, Varghese N, Schulz F, et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol*. 2020;39:499–509.
48. Bushnell B. BBMap. 2022. <https://sourceforge.net/projects/bbmap/>.
49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
51. Ecogenomics/BamM: Metagenomics-focused BAM file manipulation. <https://github.com/Ecogenomics/BamM>.
52. R Core Team. R: A language and environment for statistical computing. 2021. <https://www.r-project.org/>.
53. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIU-ViG). *Nat Biotechnol*. 2018;37:29–37.
54. Roux S, Emerson JB, Eloë-Fadrosh EA, Sullivan MB. Benchmarking viromics: an evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*. 2017;5:e3817.
55. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*. 2013;8:e61217.
56. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4:1686.
57. Ettinger C, Stajich J, Coleine C. stajichlab/Antarctic_Virus_Discovery: v2. 2022. <https://doi.org/10.5281/zenodo.7374327>.
58. Turner D, Kropinski AM, Adriaenssens EM. A roadmap for genome-based phage taxonomy. *Viruses*. 2021;13:506.
59. Selbmann L, Onofri S, Coleine C, Buzzini P, Canini F, Zucconi L. Effect of environmental parameters on biodiversity of the fungal component in lithic Antarctic communities. *Extremophiles*. 2017;21:1069–80.
60. Jarett JK, Džunková M, Schulz F, Roux S, Paez-Espino D, Eloë-Fadrosh E, et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J*. 2020;14:2527–41.
61. Coleine C, Stajich JE, Pombubpa N, Zucconi L, Onofri S, Canini F, et al. Altitude and fungal diversity influence the structure of Antarctic cryptoendolithic Bacteria communities. *Environ Microbiol Rep*. 2019;11:718–26.
62. Coleine C, Stajich JE, Zucconi L, Onofri S, Pombubpa N, Egidi E, et al. Antarctic cryptoendolithic fungal communities are highly adapted and dominated by Lecanoromycetes and Dothideomycetes. *Front Microbiol*. 2018;9:1392.
63. Li S, Guo R, Zhang Y, Li P, Chen F, Wang X, et al. A catalog of 48,425 non-redundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience*. 2022;25:104418.
64. Santos-Medellin C, Estera-Molina K, Yuan M, Pett-Ridge J, Firestone MK, Emerson JB. Spatial turnover of soil viral populations and genotypes overlap by cohesive responses to moisture in grasslands. *PNAS*. 2022;119:e2209132119.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

