nature portfolio

Peer Review File

Population ecology and biogeochemical implications of ssDNA and dsDNA viruses along a permafrost thaw gradient

Corresponding Author: Dr Gareth Trubl

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors conduct excellent viromic work on permafrost soil, which is both a technically challenging environment to study and also plays an important role in the ongoing climate crisis. Of particular note is the fact that this is one of the few studies to measure abundance of both ss and dsDNA viruses in a non-biased way (i.e. without excluding or biasing towards the detection of ssDNA). Studies of this type are rare and this work will serve as an extremely valuable benchmark for future research aiming to accurately assess viral abundance in a variety of ecosystems. As such, this work is very much suitable for publication in Nat Comm.

I do have some moderate criticism on the downstream analysis of their data that should be addressed before publication:

1) I am worried that the authors might be underusing their valuable ssDNA dataset by a) limiting their search to Microviridae and Cressdnaviricota and b) imposing contig size cutoffs for them.
Why were no searches for filamentous ssDNA phages (Loebvirae) conducted? Large numbers of them have previously been detected in soil environments (https://doi.org/10.1038/s41564-019-0510-x). In particular those might carry AMGs and other interesting cargo.
Similarly, no searches for Cossaviricota or Trapavirae (although sampling methods might preclude their detection).
Furthermore, ssDNA viruses have been found in the Varidnaviria realm, were searches conducted for those? Related, how many small complete circular contigs were detected that did not correspond to any of these groups? These contigs might be valuable when it comes to finding novel small viruses.

Searching for these additional virus groups certainly falls within the expertise of the authors – if none are found, their absence should at least be noted.

As for size cutoffs, while these are based on what is known about the biology of the viruses in question, the previous undersampling of soil viromes means that there could be surprises that the authors are missing because of that. If the authors got rid of the contig size cutoffs and instead analyzed circular contigs, would they find more divergent viruses?

In general, taxonomic annotation of the detected vOTUs would add extra value to this work. I suggest the authors at the very least run their contigs through vcontact2 or a similar taxonomy assessment program and report the findings. Linking this taxonomy to AMG content could be very useful for researchers studying individual viruses instead of viromes overall.

2) I am not entirely sure that what the authors assume to be accessory metabolic genes are actually that. In particular, I would like to point out that peptidases, glycosyl-transferases and glycosyl-hydrolases play important role in lysis and immune escape for phages. The authors should comment on the possibility of this at least.

On a nitpicky point, would it not be better to write out carbon instead of abbreviating it with C? Both from a read- and searchability perspective.

Line for line comments:

129: What does SRSLY mean?

140: Out of curiosity, why did the authors not use virsorter2?
151: pushes? I am not quite sure what is meant here.
167-170 (in line with my major comment): some rare but seemingly complete microviral MAGs <4kb https://doi.org/10.1093/ve/veac123 were recently detected, amongst other in soil environments. Similarly, there are reports of microvirus MAGs >8kb in soil/plant environments https://doi.org/10.3389/fmicb.2021.754245. While unlikely to affect the overall results, would it not be better to not have a size range cutoff and simply determine terminal redundancy for small circular ssDNA viruses?

173: coupled
176: starting
211: given how the crispr and phage datasets are separated in time, would it not make sense to look for spacers with 1bp difference - presumably virus sequences would have evolved to avoid matches? Conservatism is fine, but some value might be extracted from also looking for slightly off-target spacers (see also lines 382-385).

235: it is of course unlikely that smaller viruses (<5kb) carry amgs, but rerunning this analysis on all genomes that are at least circular could still bring up something interesting. Similarly, some taxonomic separation (see major comments) would be very valuable in this analysis.

306: What do the percentages in parenthesis mean?
328: The authors note here what is my main point of criticism. I don't see why the analysis would not be broader. How many small circular ssDNA contigs are there that do not fall in the two groups analyzed so far?
335: unclear what vOTU signal means (a bit jargony) - number of reads recruited to a specific otu? Please clarify. It is also unclear what the columns in table 7 mean - no number is higher than 100 and they dont add up to 100 so it can't be percentage? Or does it indicate how much of a given OTU is covered in a specific sample? Either way I don't really understand how it related to the sentence.

406: The author need not be too cautious with their inference – numerous studies have found microvirus prophages in Bacteroidia, which also carry abundant crispr spacers against then. As such the description of the host range of microviruses in the next sentence also needs amendments.
408: I have never heard Enterobacteria described as intracellular parasitic bacteria – everyone's favorite E. coli certainly isn't.
411: Kolton et al.'s paper does not deal with microviruses, I think the citation is misplaced
433: What does novel mean in this context?
488: Citations should be added here.
Figure 2D: What do the y-axis labels mean? Which diversity metrics are those?
Figure 3A: Extreme nitpick, but it's not a stacked bar chart if all the bars are the same length. This comes up multiple times in the figure legends.
Figure 3B: Just to clarify, with coverage the authors mean relative amount of reads mapping onto those two groups?
Supplementary Table 1: It would be nice to have more descriptive headers here.

Reviewer #2

(Remarks to the Author)
The work builds on previous work from the same group on the role of viruses in soils from northern peatlands. With a particular focus on ssDNA viruses , by the use of multiple kits. The work then goes onto look at the ecology of viruses within this environment and the potential role in biogeochemical cycling. The work is very well written and a pleasure to read.

Minor comments:
The authors have made their reads available and the genomes of ssDNA viruses. As far as I can tell the sequence of the vOTUs are not available or genes found within them. For the work to be useful to others, the entire workflow would have to be repeated to get these sequences . Please make the sequences of all vOTUs available and genes on them.

L151 – a brief explanation of why biological signal cannot be seen without 10% coverage, would help, suggesting heterogeneity is not immediately clear why. Heterogeneity at the ANI level is easier to understand , and thus dropping the 90% ANI is easier to understand. But dropping the coverage across a genome is not, without further explanation

L157 – how was ANI of 90% controlled after bowtie2 mapping ? details are not provided and bowtie2 would map reads with a lower identify than this

|L177 version and settings
L179 BBmap version and settings, what identify of reads ?
L185 – while I have sympathy with the speed of change , this is potentially 24 months ago. Specifying the month in 2021 would help here . I can see this is included later sentences
L197 versions of MAFFT, trimAl
L207 version of minced used
L208 here and elsewhere, numbers <10 should be written in full two not 2

L267 – No details are provided on the completeness of these vOTUs, which seems odd given the authors , one of which authored the go to tool to gain this information. Having an indication of the completeness of these vOTUs would give some context to the rest of the data. Are they all fragments or are any of the vOTUs HQ/complete genomes ?

L283 – tagmenation bias was noted in PMID: 27280068, which was before many of the cited references
L408 intracellular parasitic Gram-negative bacteria, is not an accurate description of Enterobacteria

L463- where the CAZyme types that are known to be linked with tail fibres and depolymerases removed from this analysis, as these are not likely to be linked to cycling ?

Reviewer #3

(Remarks to the Author)
Trubl et al. applied two library kits to recover dsDNA and ssDNA viral sequences from the viral-like particles extracted from soil samples along a permafrost thaw gradient. The studied viromes 1) are unique with few vOTUs shared with the other ecosystems and datasets; 2) assemble differently along the thaw gradient; 3) are associated with dominant microbial taxa and 4) are detected with a broad range of AMGs mostly involved in carbon metabolism. Overall, the research question is important and could be an interest to the readers of Nature Communications.

My main concerns are,
1. The manuscript contains investigations on both sequencing methods (library kit) and viral ecology along a permafrost thawing gradient. Compared to the viral ecology in permafrost that has been previously studied, the method part may be more worth exploring. If the authors agree, the ecology part can be used as a case study to highlight the high coverage that further provides opportunities for unveiling more information that has not been reported previously.
2. 'The Accel NGS 1S plus was the designation at Swift BioSciences which is no longer in business. The comparable kit for the new manufacturer (Integrated DNA technologies) is "xGen ssDNA and low-input"'. Is there any related study to justify that the alternative kit is 'comparable' to the old kit? Based on the names, it seems the old one is for both ssDNA and dsDNA detection but the alternative kit is only for ssDNA? If so, then the main concern is how reproducible this work is if it relies on a kit that is no longer available.

My minor comments,
1. Line 101 and 107, maybe 'the improvement of vOTU recovery' is partially because of more sampling efforts (20 samples in this study VS 7 viromes in Trubl et al. 2018) to recover the spatial and temporal diversity of the viral communities? If so, the last sentence in the introduction may overstate the importance of the pipeline.
2. Line 142-143, please provide the NCBI accession numbers or equivalent for the phage genomes.
3. Line 141, please provide the versions of the tools applied.
4. Line 148, what are the p values for the contigs screened by only applying VirFinder score > 0.9? p-value was computed by comparing the score with the null distribution. In addition to the absolute score, p-value is normally included to make decisions.
5. Line 149, what are the screening criteria for using CheckV?
6. Line 167-168, please clarify what source was used to curate the MCP and Rep protein sequences to build new HMMs. Please enclose the HMMs in data availability so that the others can reproduce the steps.
7. Line 178-179, were the same reads mapping thresholds applied to ssDNA viruses as specified for dsDNA viruses?
8. 'Due to the dependence of DRAM-v on VirSorter for determining confidence, the AMGs and metabolic genes discussed here were from viral contigs called by VirSorter.' The viral contigs identified by the other detection tools were not examined for AMGs? it sounds unfortunate to throw out data because of tool dependency.
9. 'To minimize the number of false positives, we considered only genes that were flanked by viral or viral-like genes on either side.' This screening criterion may be too relaxed and it is likely to introduce false positives. Screening for genes with viral or viral-like genes on both sides is recommended.
10. Line 273, Gramma check.
11. Line 278, please double check, 'significantly better' at 'p > 0.01'? a typo? Add the p-value in Fig S1C to line 279.
12. Please provide the units for each axis of plots in Fig S1.
13. It is hard to draw the conclusion of 'larger variation between the kits than variation within each kit for both thresholds' based on Fig S1D and E, especially for the part 'variation within each kit'. Maybe NMDC plots will help?
14. Line 281, 'Accel kit contigs recruited 74% more basepairs'. Is this because of a higher percent of reads assembled into viral contigs in general or a higher coverage of some of the large viral genomes?
15. The authors have compared the quantification of viral contig coverage in both library prep. How were the different quantification results from the two library prep (esp. for the dsDNA viruses that were recovered by both methods) used in the permafrost data to compare across thawing gradient/depth/year? Average? Select one over the other?
16. Line 335, perhaps it corresponds to Fig 3a not 3C? Are we missing some figures or tables to demonstrate the information in lines 337-340?
17. A clarification question, is the vOTU sharing analysis for both dsDNA and ssDNA viruses or just dsDNA viruses? It may help to make it clear if we specify the types of viruses that vOTUs represent in each analysis (i.e., vOTU sharing analysis, richness/abundance comparison).
18. Fig5 is very eye-catching but a bit hard to read to extract the info that is mentioned in the results and discussion. Maybe we could try to simplify the tree a bit (e.g., collapse some branches)?

19. Could the authors provide a sentence or two to explain how ‘Based on our phylogenic comparison of microviruses with those identified in another peatland’ can be used to ‘infer four microviruses vOTUs to infect bacteria from the phylum Bacteroidota’.
20. Line 420-422, it seems there is a logic gap between the bloom of ciliate viruses to their roles in carbon cycling. Please add a sentence or two to explain.
21. Line 425, which reference is the ‘Wu et al. 2022’? there is only one ‘Wu et al. 2022’ in the reference list that is ‘108. Wu, R., Smith, C. A., Buchko, G. W., Blaby, I. K., Paez-Espino, D., Kyrpides, N. C., Yoshikuni, Y., McDermott, J. E., Hofmockel, K. S., Cort, J. R. & Jansson, J.K. (2022). Structural characterization of a soil viral auxiliary metabolic gene product–a functional chitosanase. Nature Communications, 13(1), 1-14.’. But it is not about RNA viruses infecting Amoeba. Please check your reference list carefully.
22. Did you find any pattern in host compositions across thawing gradient/depth profile/years?
23. Line 433, AMGs that ‘were previously reported’ refer to the ones that have been reported in permafrost or generally in soil? AMGs that ‘were previously experimentally verified’ are mostly about photosynthesis?
24. Fig 7D: a nice summary of the GHs.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have taken two years to carefully address issues pointed out in the first round of reviews, including a comprehensive re-analysis of their dataset using state of the art methods. At this point, it is about time this paper is published!

I have a few minor points below:

Line-by-line
109: Optimized methods – citations need to be updated to the methods used by the authors
151: typo, citation is in the word, then citation is not upper cased
298: underlaid?
314: is amplified the right word here? To me it implies introducing bias
320: which ones are the PCR amplified metagenomes? Not specifically stated in text
333: which ones are tagmentation based? See above
631: interesting – any speculation on what the use of such a gene would be to viruses?

Reviewer #3

(Remarks to the Author)
The authors have adequately answered all my questions and I have no further comments. Nice work!

We would like to thank the editor and reviewers for the invitation to revise our manuscript and for the positive and constructive feedback. We believe we have fully addressed all points raised.

Reviewer #1

The authors conduct excellent viromic work on permafrost soil, which is both a technically challenging environment to study and also plays an important role in the ongoing climate crisis. Of particular note is the fact that this is one of the few studies to measure abundance of both ss and dsDNA viruses in a non-biased way (i.e. without excluding or biasing towards the detection of ssDNA). Studies of this type are rare and this work will serve as an extremely valuable benchmark for future research aiming to accurately assess viral abundance in a variety of ecosystems. As such, this work is very much suitable for publication in Nat Comm.

Thank you for taking the time to review our work, these kind words, and your assessment of the specific contribution this work makes to the available literature.

I do have some moderate criticism on the downstream analysis of their data that should be addressed before publication:

1) I am worried that the authors might be underusing their valuable ssDNA dataset by a) limiting their search to Microviridae and Cressdnaviricota and b) imposing contig size cutoffs for them. Why were no searches for filamentous ssDNA phages (Loebvirae) conducted? Large numbers of them have previously been detected in soil environments (https://doi.org/10.1038/s41564-019-0510-x). In particular those might carry AMGs and other interesting cargo.
Similarly, no searches for Cossaviricota or Trapavirae (although sampling methods might preclude their detection). Furthermore, ssDNA viruses have been found in the Varidnaviria realm, were searches conducted for those? Related, how many small complete circular contigs were detected that did not correspond to any of these groups? These contigs might be valuable when it comes to finding novel small viruses.

Searching for these additional virus groups certainly falls within the expertise of the authors – if none are found, their absence should at least be noted.

As for size cutoffs, while these are based on what is known about the biology of the viruses in question, the previous undersampling of soil viromes means that there could be surprises that the authors are missing because of that. If the authors got rid of the contig size cutoffs and instead analyzed circular contigs, would they find more divergent viruses?
Thank you for the excellent suggestions. While the manuscript was in review, a new detection tool called geNomad (Camargo et al., 2023) was released, which uses marker genes to detect viruses. To address your concerns, we ran the contigs with no size cutoffs through geNomad to

get a marker-gene based taxonomy (includes the marker genes from Roux et al., 2019 for Inoviridae as you suggested), which should flag any additional ssDNA viruses we may not have discussed. For vOTU consideration, we retained ssDNA viral contigs >500 bases and then clustered the viral contigs using the same thresholds used for the dsDNA vOTUs (95% average nucleotide identity and 85% alignment fraction). We have now included our new findings in Supplementary Table 6. In the end, we identified 54 new ssDNA vOTUs, 19 in the order Tubulavirales, 35 in the order Piccovirales. Additionally, (1) all our circular microviruses range from 3,962–5,784 bases, (2) there are 38 Loebvirae (inoviruses) in the size range 256–7,861 bases that are not circular, (3) 1,462 microviruses 246–7213 bases size range with 162 being circular.

Camargo, A.P., Roux, S., Schulz, F. et al. Identification of mobile genetic elements with geNomad. Nat Biotechnol (2023). https://doi.org/10.1038/s41587-023-01953-y

In general, taxonomic annotation of the detected vOTUs would add extra value to this work. I suggest the authors at the very least run their contigs through vcontact2 or a similar taxonomy assessment program and report the findings. Linking this taxonomy to AMG content could be very useful for researchers studying individual viruses instead of viromes overall.

Thank you for the suggestion. We ran the dsDNA vOTUs through vcontact2 to get genus-level taxonomic clustering results (Table S4).

These data were incorporated into the manuscript using the following text: “First the 9,963 dsDNA vOTUs were compared to a set of virus reference genomes (viral refseq v.85) using a gene-sharing network as implemented in vConTACT2 (v. 0.9.19; Jang et al. 201954). This tool clustered 5,261 of the 9,663 vOTUs into 1,488 genus-level viral clusters (VCs) and only 34 (~2.3%) of the VCs included a reference genome (Table S4), highlighting the novelty of our dataset.”

2) I am not entirely sure that what the authors assume to be accessory metabolic genes are actually that. In particular, I would like to point out that peptidases, glycosyl-transferases and glycosyl-hydrolases play important roles in lysis and immune escape for phages. The authors should comment on the possibility of this at least.

To the first question, we assigned AMGs using the standard approaches that have emerged from several international workshop discussions and are summarized in a recent publication (Pratama et al., 2021). Regarding the second part of your question on characterization of AMGs, thank you for raising this point because we could be clearer. We only went into detail for the glycoside hydrolases that are known to breakdown complex carbohydrates. We have now updated our supplementary table 10, to only include AMGs and not other metabolic genes. This has removed all peptidases and glycosyl-transferases. Additionally, we updated our analyses to search for AMGs in on new vOTU dataset from VirSorter2 and geNomad through a recently

developed modular viromics pipeline (v 1.0; Coclet et al. 2024) designed to generate the required affiliation file format compatible with DRAM-v. There are now 811 putative AMGs. We define AMGs as genes with a DRAM-v annotated score from 1 to 3 (1 being most confident and 5 being no confidence; Shaffer et al. 2020) and either a metabolic gene in a KEGG metabolic pathway (class 1) or general metabolic function or not involved in a KEGG metabolic pathway (class 2) as previously suggested (Hurwitz, Brum & Sullivan 2015). To minimize the number of false positives, we considered only genes that were flanked by viral or viral-like genes on both sides, and removed genes with a 'viral' flag (V; assigned when the gene has been associated with a VOGDB identifier with the replication or structure categories), transposon flag (T; given when the gene is on a contig that contains a transposon), attachment flag (A; assigned when the gene, has been associated with viral host attachment and entry as is the case with many CAZymes), or viral-like peptidase flag (P; is similar to the "A" flag but when the gene is given identifiers that are peptidases previously identified as potentially-viral using, not AMGs, based on the distribution of peptidase families provided in the MEROPS database) or were removed because the gene did not have a gene ID or a gene description.

We added the following text to the manuscript to be clearer "Identification and curation of AMGs was performed following a recently established standard operating protocol (Pratama et al. 2021). First, all detected viruses were annotated with DRAM v 1.0 (Table S9) and then AMGs were detected for vOTUs with DRAM-v. Because DRAM-v relies on outputs VirSorter/Virsorter2 for determining confidence in classifying AMGs, we ran the geNomad-identified vOTUs through a recently developed modular viromics pipeline (v 1.0; Coclet et al. 2024) designed to generate the required affiliation file format compatible with DRAM-v. We then identified AMGs using DRAM-v with the --use_uniref flag enabled and the following databases: PFAM v33.1, KEGG v89.1, dbCAN v9, MEROPS v120, and VOGDB database (Dec. 2024) (Shaffer et al. 2020). For each gene on a viral contig that DRAM-v annotated as metabolic (M flag), a score, from 1 to 5 (1 being most confident), denotes the likelihood that the gene belongs to a viral genome rather than a degraded prophage region or a poorly defined viral genome boundary (Shaffer et al. 2020). To link these viruses more readily to the carbon cycle in this system, we included 46 genes as putative AMGs (added to DRAM database) that were highlighted to be important in ecosystem carbon cycling (Woodcroft et al. 20187; Table S10). To minimize the number of false positives, we considered only genes that had a gene ID or a gene description and removed genes with a 'viral' flag (V), transposon flag (T), viral-like peptidase (P) or attachment flag (A). If the AMG had been previously identified in the literature according to the DRAM-v database, it received a "K" flag (known) and if the AMG was known and had also been experimentally verified in a previous study (meaning it has been shown in a host to provide a specific function), it received an "E" flag and was labeled "experimentally verified". The remaining AMGs were labeled as "putative AMGs". For category assignment (e.g., carbon utilization), AMGs were assigned first by DRAM-v provided categories, and if no category was provided by DRAM-v, the genes were manually assigned to each category based on DRAM annotation."

Pratama, A.A., Bolduc, B., Zayed, A.A., Zhong, Z., Guo, J., Vik, D., Gazitúa, M.C., Wainaina, J., Roux, S. & Sullivan, M.B. 2021. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. PeerJ (9):e11447.

Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, S.L., Solden, L.M., Liu, P., Narrowe, A.B., Rodríguez-Ramos, J., Bolduc, B. and Gazitúa, M.C., 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic acids research*, *48*(16), pp.8883-8900.

Hurwitz, B.L., Brum, J.R. and Sullivan, M.B., 2015. Depth-stratified functional and taxonomic niche specialization in the 'core'and 'flexible'Pacific Ocean Virome. The ISME Journal, 9(2), pp.472-484.

Coclet, C., Camargo, A.P. and Roux, S., 2024. MVP: a modular viromics pipeline to identify, filter, cluster, annotate, and bin viruses from metagenomes. *Msystems*, *9*(10), pp.e00888-24.

On a nitpicky point, would it not be better to write out carbon instead of abbreviating it with C? Both from a read- and searchability perspective.
Done.

Line for line comments:

129: What does SRSLY mean? This is the name of the kits by ClaretBio (https://www.claretbio.com/products/kits) that can quantitatively amplify ssDNA and dsDNA viruses. We rephrased the writing to be clearer "Recently, the SRSLY kits (ClaretBio, Santa Cruz, CA) were proposed and could provide another option for quantitative ssDNA/dsDNA viromics.".

140: Out of curiosity, why did the authors not use virsorter2?
The read QC, assembly, and virus detection was done in 2018 before VirSorter2 was published. As we were finishing up the other analyses the pandemic hit. We agreed that this dataset would benefit from running the contigs through VirSorter2. In addition, we used geNomad to help with detection of ssDNA viruses. Our vOTU dataset increased from 9,560 to 9,963. All of our previous vOTUs were detected, but additionally, we captured 54 new ssDNA vOTUs and 349 dsDNA vOTUs.

151: pushes? I am not quite sure what is meant here.
To clarify, we used "pushes" to mean push the limit, stress, strain. We have replaced pushes with strains to clarify. The sentence of the manuscript now reads "As is common for soil virome work, only a small amount of DNA was extracted from these samples which strains the library

preparation methods and requires relaxation of read-mapping along the contigs requirements to see 'biological' signals (Páez-Espino et al. 2016; Roux et al. 2017; Trubl et al. 2019).".

167-170 (in line with my major comment): some rare but seemingly complete microviral MAGs <4kb https://doi.org/10.1093/ve/veac123 were recently detected, amongst other in soil environments. Similarly, there are reports of microvirus MAGs >8kb in soil/plant environments https://doi.org/10.3389/fmicb.2021.754245. While unlikely to affect the overall results, would it not be better to not have a size range cutoff and simply determine terminal redundancy for small circular ssDNA viruses?
Thank you for the suggestion, which sounds like a better option for the smaller ssDNA viruses. To address this, we have added new marker genes for ssDNA virus detection using the new tool recently available (geNomad) and this time running all contigs (i.e., no size limits). New ssDNA viruses detected were kept if >500 bases and then clustered as previously stated (95% ANI and 85% AF). No additional microvirus or cressdnaviruses vOTUs were identified.

173: coupled
Done.

176: starting
Done.

211: given how the crispr and phage datasets are separated in time, would it not make sense to look for spacers with 1bp difference - presumably virus sequences would have evolved to avoid matches? Conservatism is fine, but some value might be extracted from also looking for slightly off-target spacers (see also lines 382-385).
Absolutely and we used a conservative approach since the iPHoP host predictor pipeline (the other approach we used) does a relaxed approach. It first performs a CRISPR array analysis considering only hits to spacers 25 nucleotides or longer, with less than 8 mismatches overall, and with a custom complexity score <0.6). Then the CRISPR spacers are filtered to only keep the hits with 4 or less mismatches for the machine-learning classifier. The iPHoP output provides a confidence score and includes which methods it used for host prediction. For matches where CRISPR spacer was the main method it is listed in column E "List of methods" and for a CRISPR spacer to be the only virus-host linkage, then it has to have 2 mismatches or less.

235: it is of course unlikely that smaller viruses (<5kb) carry amgs, but rerunning this analysis on all genomes that are at least circular could still bring up something interesting. Similarly, some taxonomic separation (see major comments) would be very valuable in this analysis.

Thank you for the suggestion. We have now searched all vOTUs for AMGs. First, we used DRAM to annotate all vOTUs (now provided in Supplementary Table 10). We then updated our AMG analyses using the new vOTU dataset derived from VirSorter2 and geNomad, processed through

a recently developed modular viromics pipeline (v1.0; Coclet et al., 2024) that generates DRAM-v-compatible affiliation files. This analysis identified one ssDNA vOTU, an unclassified cressdnavirus (F_XD_2016_NODE_2697), carrying a single AMG annotated as 6-pyruvoyl tetrahydropterin synthase. This enzyme is involved in cofactor biosynthesis and may support carbon or nitrogen metabolism under stress or competitive conditions. This viral AMG was previously reported in Roux et al., 2016.

For the second part of your question regarding taxonomy, we updated ssDNA taxonomy using geNomad and we ran the dsDNA vOTUs through vcontact2 to get genus-level taxonomic clustering results (Table S4).

These data were incorporated into the manuscript using the following text: "First the 9,963 dsDNA vOTUs were compared to a set of virus reference genomes (viral refseq v.85) using a gene-sharing network as implemented in vConTACT2 (v. 0.9.19; Jang et al. 2019). This tool clustered 5,261 of the 9,663 vOTUs into 1,488 genus-level viral clusters (VCs) and only 34 (~2.3%) of the VCs included a reference genome (Table S4), highlighting the novelty of our dataset."

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J. and Pesant, S., 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature, 537(7622), pp.689-693.

306: What do the percentages in parenthesis mean?

It was just another way to represent the data. Since the number of vOTUs is there, we have removed it.

328: The authors note here what is my main point of criticism. I don't see why the analysis would not be broader. How many small circular ssDNA contigs are there that do not fall in the two groups analyzed so far?

Thank you for the suggestion. To address this, we have added new marker genes for ssDNA virus detection using the new tool recently made available during review (geNomad) and this time running all contigs (i.e., no size limits). New ssDNA viruses detected were kept if >500 bases and then clustered as previously stated (95% ANI and 85% AF). While no additional microvirus or cressdnaviruses vOTUs were identified, this updated and more permissive analysis did identify 54 new ssDNA vOTUs, 19 in the order Tubulavirales, 35 in the order Piccovirales. All of these data are now included in the manuscript, and we appreciate the reviewer pushing us to maximize detecting the ssDNA virus signal in these datasets so that these genomic data can be more readily available for follow-on studies.

335: unclear what vOTU signal means (a bit jargony) - number of reads recruited to a specific otu? Please clarify. It is also unclear what the columns in table 7 mean - no number is higher than 100 and they dont add up to 100 so it can't be percentage? Or does it indicate how much

of a given OTU is covered in a specific sample? Either way I don't really understand how it related to the sentence.
The percentage in a cell is how much of the vOTU genome is covered by the mapped reads. For example, If the reads map to two regions of the same genome in one area covering 25% and another area covering 40% of the genome, then the total coverage is 65%. Since this was confusing and we had to redo the coverage for the new ssDNA vOTUs, we removed these data and the associated text and provided relative abundance data using the 75% strict threshold that we discuss in the methods for the dsDNA vOTUs.

406: The author need not be too cautious with their inference – numerous studies have found microvirus prophages in Bacteroidia, which also carry abundant crispr spacers against then. As such the description of the host range of microviruses in the next sentence also needs amendments.
Great suggestions, thank you. We have removed the word cautious since we already say inferred and updated the host range to explicitly mention Bacteroidia. The sentence in the revised manuscript now reads "Based on our phylogenic comparison of microviruses with those identified in another peatland, Les Pradeaux Mire (Fig. S3) (Quaiser et al., 2015), we infer four microviruses vOTUs to infect bacteria from the phylum Bacteroidota.".

408: I have never heard Enterobacteria described as intracellular parasitic bacteria – everyone's favorite E. coli certainly isn't.
Our apologies, that phrasing does not describe all individuals of that group and it has been removed. There are members Enterobacteriaceae such as Klebsiella, Shigella, and Salmonella that have been described this way in the literature, but this description can have multiple meanings and they do not apply here (see citations below). To make this clearer, we have revised our wording and the text now simply states "Gram-negative bacteria".

-Clancy, C. F. (2019). Enterobacteriaceae. In Handbook of Microbiology (pp. 221-229). CRC Press.
-Beyene, G., & Tasew, H. (2014). Prevalence of intestinal parasite, Shigella and Salmonella species among diarrheal children in Jimma health center, Jimma southwest Ethiopia: a cross sectional study. Annals of clinical microbiology and antimicrobials, 13, 1-7.
-Pattison, M., McMullin, P., Bradbury, J. M., & Alexander, D. (Eds.). (2007). Poultry diseases. Elsevier Health Sciences.
-Linton, A. H., & Hinton, M. H. (1988, January). Enterobacteriaceae associated with animals in health and disease. In Society for Applied Bacteriology symposium series (Vol. 17, pp. 71S-85S)

411: Kolton et al.'s paper does not deal with microviruses, I think the citation is misplaced
Thank you for flagging that issue, it was a mistake. The misplaced reference has now been removed.

433: What does novel mean in this context?

To clarify, we used novel to describe that this is the first identification of these genes to be identified in a virus with this function based on the DRAM-v database. We have removed the word novel in the revised manuscript and the text now reads “We identified 811 putative AMGs from 658 vOTUs, 652 AMGs were previously reported in viruses (see methods) and 30 of which were previously experimentally verified, i.e., shown in a microbial host to provide a specific function (Fig. 7A).”

We also clarified in the methods sections with the following text “If the AMG had been previously identified in the literature according to the DRAM-v database, it received a “K” flag (known) and if the AMG was known and had also been experimentally verified in a previous study (meaning it has been shown in a host to provide a specific function), it received an “E” flag and was labeled “experimentally verified”. The remaining AMGs were labeled as “putative AMGs”.”

488: Citations should be added here.
Thank you for flagging this. The following citations were added: Woodcroft et al., 2018 for microbial diversity and Hough et al., 2022 for plant productivity.

Figure 2D: What do the y-axis labels mean? Which diversity metrics are those?
Our apologies, the abbreviations have been removed and replaced by the words (i.e., richness, Shannon–Wiener diversity index, and Pielou's evenness).

Figure 3A: Extreme nitpick, but it’s not a stacked bar chart if all the bars are the same length. This comes up multiple times in the figure legends.
We appreciate the input and have change the description to “proportional bar chart”.

Figure 3B: Just to clarify, with coverage the authors mean relative amount of reads mapping onto those two groups?
Yes and throughout the manuscript, we have changed coverage to relative abundance. The text has been revised to be clearer “(A) A breakdown of ssDNA vOTUs at the phylum taxonomic level. (B) Family-level taxonomic resolution of vOTUs classified within Cressdnaviricota, Phixviricota, Cossaviricota, and Hofneiviricota. (C) Habitat-specific relative abundance of taxonomically classified ssDNA vOTUs across palsa, bog, and fen habitats. The number at the end of each bar represents the number of vOTUs detected.”

Supplementary Table 1: It would be nice to have more descriptive headers here.
Table S1 has been updated to be clearer, such as abbreviations being removed and columns without data being removed. Text now reads “(A) Total number of viral contigs recovered per sample, showing consistent recovery across library methods. Read counts mapping to contigs detected by both kits at a relaxed (B) threshold and strict (C) threshold. NMDS plots of Bray-Curtis dissimilarities based on contig abundance, comparing library preparation methods at the

relaxed (D) and strict (E) mapping thresholds. Samples cluster by individual rather than by library method, and PERMANOVA and ANOSIM tests indicate no significant differences between kits (PERMANOVA $p > 0.88$; ANOSIM $R < 0.02$, $p > 0.75$)."

Reviewer #2 (Remarks to the Author):

The work builds on previous work from the same group on the role of viruses in soils from northern peatlands. With a particular focus on ssDNA viruses , by the use of multiple kits. The work then goes onto look at the ecology of viruses within this environment and the potential role in biogeochemical cycling. The work is very well written and a pleasure to read.
Thank you for your review, and these kind words.

Minor comments:
The authors have made their reads available and the genomes of ssDNA viruses. As far as I can tell the sequence of the vOTUs are not available or genes found within them. For the work to be useful to others, the entire workflow would have to be repeated to get these sequences. Please make the sequences of all vOTUs available and genes on them.
Our apologies and thank you for the suggestion. We now make a broader suite of data products available with the sequences for the vOTUs available at https://doi.org/10.5281/zenodo.10251898 and the predicted genes at 10.5281/zenodo.8384701. In the revised manuscript, this information is made clear under the data availability section.

L151 – a brief explanation of why biological signal cannot be seen without 10% coverage, would help, suggesting heterogeneity is not immediately clear why. Heterogeneity at the ANI level is easier to understand, and thus dropping the 90% ANI is easier to understand. But dropping the coverage across a genome is not, without further explanation

Thank you for the helpful comment. To clarify, the 10% coverage threshold was only applied during our comparison of library preparation methods, not in the final analysis of viral populations. We selected this relaxed threshold to maximize read recruitment and better detect low-abundance viral signals that may differ between library kits. While this threshold increases the likelihood of including partially sampled viral populations, we acknowledge it comes with decreased confidence in population-level quantification. To maintain rigor, all primary ecological and functional analyses were conducted using the more stringent and widely adopted 75% coverage threshold (e.g., Roux et al., 2017). The relaxed 10% threshold was used solely as a comparative exploratory tool to assess the sensitivity of each library method.

We have removed the term "heterogeneity" to avoid confusion and clarified this distinction in the revised methods:

"Therefore, reads were mapped back to the nonredundant set of contigs at 90% ANI and two thresholds for the length of contig covered with Read2RefMapper v1.1.1 (Bolduc et al., 2017) to estimate coverage: 75% contig coverage for our strict threshold as it offers robust coverage of well-sampled populations (Roux et al. 201757), and 10% for our relaxed threshold because it increases read recruitment to the contigs which we hypothesized would allow detection of less abundant viruses and highlight more differences between the library kits (Páez-Espino et al. 2016; Páez-Espino et al. 2017; Trubl et al. 2019). The relaxed coverage threshold was only used for library comparisons and the strict threshold was used for characterization of the vOTUs."

L157 – how was ANI of 90% controlled after bowtie2 mapping ? details are not provided and bowtie2 would map reads with a lower identify than this

Thank you for raising this point. The tool that was used was Read2RefMapper (Bolduc et al., 2017).

We added a link under Data availability (https://bitbucket.org/bolduc/docker-read2refmapper/src/master/) and updated the following text in the methods section "Therefore, reads were mapped back to the nonredundant set of contigs at 90% ANI and two thresholds for the length of contig covered with Read2RefMapper v1.1.1 (Bolduc et al., 2017)".

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L., & Sullivan, M. B. (2017). iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *The ISME journal*, *11*(1), 7-14.

L177 version and settings

Thank you. "VirSorter v1.0.3; virome decontamination mode" was added. We have since used an updated version of VirSorter and re-run our analyses. The version was cited "VirSorter2 v2.2.4"

L179 BBmap version and settings, what identify of reads ?

BBMap v 38.94. Run with default setting coupled with nullifybrokenquality flag, and 1-2 nucleotide variation per read.

L185 – while I have sympathy with the speed of change , this is potentially 24 months ago. Specifying the month in 2021 would help here . I can see this is included later sentences

Thank you for your understanding and we added "May" to specify the month.

L197 versions of MAFFT, trimAl

MAFFT v7.505 and trimAl v1.3 were added. Thanks.

L207 version of minced used

Version 0.3.2 was added to the text.

L208 here and elsewhere, numbers <10 should be written in full two not 2
This was corrected throughout the manuscript, except where the number symbolizes set criteria for a method or figure/table (e.g., AMGs scores 1-3).

L267 – No details are provided on the completeness of these vOTUs, which seems odd given the authors , one of which authored the go to tool to gain this information. Having an indication of the completeness of these vOTUs would give some context to the rest of the data. Are they all fragments or are any of the vOTUs HQ/complete genomes ?
The information on Completedness of the vOTUs is a few lines lower and was performed with Checkv and the data was present in column "R" of Table S3. In the submitted manuscript, this is the relevant text: "We next assessed the quality of the dsDNA vOTUs with CheckV and identified 20 complete genomes, plus 716 high quality, 1,110 medium quality, 5,985 low quality, and 1,053 quality-not-determined viral genomes (Supplementary Table 3). Because CheckV performs best where closely related reference genomes are available and soil virus reference genomes are not widely available, we interpret the large number of 'not-determined' and 'low quality' dsDNA vOTU genomes to represent an unknown mix of challenges for CheckV and highly fragmented genomes from low coverage soil viromes."

L283 – tagmenation bias was noted in PMID: 27280068, which was before many of the cited references
Thank you for the suggested reference. It has been added to the text along with another early reference "Contig coverage biases, specifically unmapped reads, are a common problem with tagmentation (Marine et al. 2011; Solonenko et al. 2013b; Jones et al. 2015; Rihtman et al., 2016; Sato et al. 2019; Gunasekera et al. 2021)."

L408 intracellular parasitic Gram-negative bacteria, is not an accurate description of Enterobacteria
Our apologies, you are correct that it does not apply to all individuals, and it has been removed. There are members Enterobacteriaceae such as Klebsiella, Shigella, and Salmonella that have been described this way in the literature, but this description can have multiple meanings and they do not apply here (see below). The text now simply states "Gram-negative bacteria".

-Clancy, C. F. (2019). Enterobacteriaceae. In Handbook of Microbiology (pp. 221-229). CRC Press.
-Beyene, G., & Tasew, H. (2014). Prevalence of intestinal parasite, Shigella and Salmonella species among diarrheal children in Jimma health center, Jimma southwest Ethiopia: a cross sectional study. Annals of clinical microbiology and antimicrobials, 13, 1-7.
-Pattison, M., McMullin, P., Bradbury, J. M., & Alexander, D. (Eds.). (2007). Poultry diseases. Elsevier Health Sciences.
-Linton, A. H., & Hinton, M. H. (1988, January). Enterobacteriaceae associated with animals in health and disease. In Society for Applied Bacteriology symposium series (Vol. 17, pp. 71S-85S)

L463- where the CAZyme types that are known to be linked with tail fibres and depolymerases removed from this analysis, as these are not likely to be linked to cycling ?
Thank you for asking. The AMGs were detected and characterized with DRAM-v. In the output from DRAM-v, a 'viral' flag (V), transposon flag (T), viral-like peptidase (P) or attachment flag (A) is assigned when the gene has been associated with a VOGDB identifier with the replication or structure categories. Additionally, we manually reviewed the AMGs as outlined in Pratama et al., (2021), which is why we included the term "other metabolic genes" in the first submission to label metabolic genes used in viral infection. To be clearer, we have further classified AMGs into types I and II as first suggested in Hurwitz, Brum & Sullivan (2015) and removed metabolic gene in the new first column of the AMG table (Table S9). This is also reflected in the updated Figure 7.

Pratama, A. A., Bolduc, B., Zayed, A. A., Zhong, Z. P., Guo, J., Vik, D. R., ... & Sullivan, M. B. (2021). Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. *PeerJ*, *9*, e11447.
Hurwitz, B. L., Brum, J. R., & Sullivan, M. B. (2015). Depth-stratified functional and taxonomic niche specialization in the 'core'and 'flexible'Pacific Ocean Virome. *The ISME Journal*, *9*(2), 472-484.

Reviewer #3 (Remarks to the Author):

Trubl et al. applied two library kits to recover dsDNA and ssDNA viral sequences from the viral-like particles extracted from soil samples along a permafrost thaw gradient. The studied viromes 1) are unique with few vOTUs shared with the other ecosystems and datasets; 2) assemble differently along the thaw gradient; 3) are associated with dominant microbial taxa and 4) are detected with a broad range of AMGs mostly involved in carbon metabolism. Overall, the research question is important and could be an interest to the readers of Nature Communications.
Thank you for your review.

My main concerns are,
1. The manuscript contains investigations on both sequencing methods (library kit) and viral ecology along a permafrost thawing gradient. Compared to the viral ecology in permafrost that has been previously studied, the method part may be more worth exploring. If the authors agree, the ecology part can be used as a case study to highlight the high coverage that further provides opportunities for unveiling more information that has not been reported previously.

We agree with the reviewer that the new methodology described here is exciting. We envision a more detailed analysis of this methodological aspect but believe it would be best as a separate

manuscript rather than in this paper for clarity; as we had done for our read QC and assembly pipeline (Roux et al., 2019). Nevertheless, we added the following sentences to the conclusion to bring this point more to reader's attention: "This increased recovery of soil viral diversity through a modified library preparation protocol and read quality-control bioinformatics pipeline already provided novel insights into under-studied viral groups. Additionally, it is a step closer to an integrated strategy that would enable more comprehensive surveys of soil viruses using meta-omics, a goal the field has been working on for the past ten years."

Roux, S., Trubl, G., Goudeau, D., Nath, N., Couradeau, E., Ahlgren, N.A., Zhan, Y., Marsan, D., Chen, F., Fuhrman, J.A. and Northen, T.R., 2019. Optimizing de novo genome assembly from PCR-amplified metagenomes. PeerJ, 7, p.e6902.

2. 'The Accel NGS 1S plus was the designation at Swift BioSciences which is no longer in business. The comparable kit for the new manufacturer (Integrated DNA technologies) is "xGen ssDNA and low-input"'. Is there any related study to justify that the alternative kit is 'comparable' to the old kit? Based on the names, it seems the old one is for both ssDNA and dsDNA detection but the alternative kit is only for ssDNA? If so, then the main concern is how reproducible this work is if it relies on a kit that is no longer available.

Thank you for raising this concern. We would like to clarify that the "xGen ssDNA and low-input" kit is the same kit as Accel NGS 1S Plus, per the manufacturer. Integrated DNA technologies was the company that bought Swift biosciences which makes the Accel NGS 1S plus kit. See this news article for more: https://www.idtdna.com/pages/about/news/2021/03/11/idt-acquires-swift-biosciences-a-pioneer-in-the-development-of-next-generation-sequencing-(ngs)-library-preparation-genomics-kits-for-academic-translational-and-clinical-research#:~:text=CORALVILLE%2C%20Iowa%20(March%2011%2C,for%20academic%2C%20translational%2C%20and%20clinical.

We wanted to support the community and show an additional set of kits "SRSLY" by ClaretBio that claim to also work on both ssDNA and dsDNA. The Joint Genome Institute has tested this kit and they are working on a publication for it that includes extensive benchmarking to help guide the research community in making decisions about kits.

My minor comments,

1. Line 101 and 107, maybe 'the improvement of vOTU recovery' is partially because of more sampling efforts (20 samples in this study VS 7 viromes in Trubl et al. 2018) to recover the spatial and temporal diversity of the viral communities? If so, the last sentence in the introduction may overstate the importance of the pipeline.

Thank you for pointing this out. We have rephrased this sentence as "The combination of a larger virome dataset and our optimized sample-to-ecology pipeline had a 10-fold improvement

of vOTU recovery (~30 vOTUs/Gbp of virome versus ~3 vOTUs/Gbp of virome) compared to our previous method (Trubl et al. 2018)."

2. Line 142-143, please provide the NCBI accession numbers or equivalent for the phage genomes.
We added the RefSeq identifiers: Enterobacteria phage PhiX17, Alpha3, M13, Cellulophaga baltica phages, and Pseudoalteromonas phages, RefSeq genomes identifiers NC_001422.1, NC_001330.1, NC_003287.2, NC_021788.1, NC_021798.1, NC_021806.1, NC_021789.1, NC_021792.1, NC_021796.1, NC_021803.1, NC_021794.1, NC_021799.1, NC_021802.1, NC_021790.1, NC_021791.1, NC_021795.1, NC_021800.1, NC_021804.1, NC_021793.1, NC_021805.1, NC_021797.1, NC_048630.1, NC_070825.1, NC_070826.1, NC_070820.1, and NC_015293.1

3. Line 141, please provide the versions of the tools applied.
Thank you. VirSorter v1.0.3 virome decontamination mode (now updated to VirSorter2 v2.2.4), MAFFT v7.505, trimAl v1.3, and minced Version 0.3.2 were added to the text.

4. Line 148, what are the p values for the contigs screened by only applying VirFinder score > 0.9? p-value was computed by comparing the score with the null distribution. In addition to the absolute score, p-value is normally included to make decisions.
Thank you and sorry for not including this. We had not added the p value in the text "VirFinder ≥10kb with a score ≥0.9 and p >0.05". The text has been revised to now include VirSorter2 and geNomad.

5. Line 149, what are the screening criteria for using CheckV?
We used default parameters and included that in the text.

6. Line 167-168, please clarify what source was used to curate the MCP and Rep protein sequences to build new HMMs. Please enclose the HMMs in data availability so that the others can reproduce the steps.
We added "hmmsearch v3 (Eddy, 2009; cutoffs: score ≥ 50 and e-value ≤ 0.001)". The HMMs have been made available via Zenodo: https://doi.org/10.5281/zenodo.8353598, which was included under data availability.

7. Line 178-179, were the same reads mapping thresholds applied to ssDNA viruses as specified for dsDNA viruses?
Yes. We and we included new text to clarify "All these ssDNA viral genomes were then clustered into vOTUs and normalized coverage profiles were establish following the methods stated above for the dsDNA vOTUs."

8. 'Due to the dependence of DRAM-v on VirSorter for determining confidence, the AMGs and

metabolic genes discussed here were from viral contigs called by VirSorter.' The viral contigs identified by the other detection tools were not examined for AMGs? it sounds unfortunate to throw out data because of tool dependency.

Thank you for the suggestion. We have now searched all vOTUs for AMGs. First, we used DRAM to obtain gene annotations for all vOTUs (now Supplementary Table 9). For the vOTUs not detected by VirSorter or <10kb, we did a manual AMG curation following the guidelines outlined previously (Pratama et al. 2021).

We updated the text as "Identification and curation of AMGs was performed following a recently established standard operating protocol (Pratama et al. 2021). First, all detected viruses were annotated with DRAM v 1.0 (Table S9) and then AMGs were detected for vOTUs with DRAM-v (Shaffer et al., 2020). Because DRAM-v relies on outputs VirSorter/Virsorter2 for determining confidence in classifying AMGs, we ran the geNomad-identified vOTUs through a recently developed modular viromics pipeline (v 1.0; Coclet et al. 2024) designed to generate the required affiliation file format compatible with DRAM-v. We then identified AMGs using DRAM-v with the --use_uniref flag enabled and the following databases: PFAM v33.1, KEGG v89.1, dbCAN v9, MEROPS v120, and VOGDB database (Dec. 2024) (Shaffer et al. 2020). For each gene on a viral contig that DRAM-v annotated as metabolic (M flag), a score, from 1 to 5 (1 being most confident), denotes the likelihood that the gene belongs to a viral genome rather than a degraded prophage region or a poorly defined viral genome boundary (Shaffer et al. 2020). To link these viruses more readily to the carbon cycle in this system, we included 46 genes as putative AMGs (added to DRAM database) that were highlighted to be important in ecosystem carbon cycling (Woodcroft et al. 201; Table S10). To minimize the number of false positives, we considered only genes that had a gene ID or a gene description and removed genes with a 'viral' flag (V), transposon flag (T), viral-like peptidase (P) or attachment flag (A). If the AMG had been previously identified in the literature according to the DRAM-v database, it received a "K" flag (known) and if the AMG was known and had also been experimentally verified in a previous study (meaning it has been shown in a host to provide a specific function), it received an "E" flag and was labeled "experimentally verified". The remaining AMGs were labeled as "putative AMGs". For category assignment (e.g., carbon utilization), AMGs were assigned first by DRAM-v provided categories, and if no category was provided by DRAM-v, the genes were manually assigned to each category based on DRAM annotation."

9. 'To minimize the number of false positives, we considered only genes that were flanked by viral or viral-like genes on either side.' This screening criterion may be too relaxed and it is likely to introduce false positives. Screening for genes with viral or viral-like genes on both sides is recommended.

You are correct and that is what we did, but we see our poor word choice. We have replaced "either" with "both". Text now appears as (lines 262-263) "To minimize the number of false

positives, we considered only genes that were flanked by viral or viral-like genes on both sides." Thank you for helping us be clearer.

10. Line 273, Gramma check.

We have replaced "further" with "characterize". The text now reads "Here, we leveraged the larger dataset of paired soil viromes to characterize these two library preparation methods in terms of viral contig recovery and potential coverage biases."

11. Line 278, please double check, 'significantly better' at 'p > 0.01'? a typo? Add the p-value in Fig S1C to line 279.

Thank you. The text has now been updated as "Comparing the coverage of contigs detected in both kits for each metagenome, revealed the Accel kit performed significantly better at the relaxed threshold (Supplementary Fig. 1B; ANOVA p <0.01) and the kits performed similarly at the strict threshold (Supplementary Fig. 1C; ANOVA p 0.27)."

12. Please provide the units for each axis of plots in Fig S1.
Done.

13. It is hard to draw the conclusion of 'larger variation between the kits than variation within each kit for both thresholds' based on Fig S1D and E, especially for the part 'variation within each kit'. Maybe NMDC plots will help?
We appreciate the suggestion and changed panels D and E to NMDS plots as suggested. We have updated Figure S1 as suggested in the previous comment to make the panels clearer and added new text to the figure legend. We hope this added explanation helps clarify our logic in the current work and thank you for helping us identify the concern.

14. Line 281, 'Accel kit contigs recruited 74% more basepairs'. Is this because of a higher percent of reads assembled into viral contigs in general or a higher coverage of some of the large viral genomes?
Thank you for asking. This assessment was done using normalized reads to prevent bias from differences in sequencing depth and contig length (see lines 177-179 "Coverage was calculated as the number of bases mapped to each read normalized by the length of the contig, and by the total number of bases sequenced in the metagenome using Bowtie 2".).

15. The authors have compared the quantification of viral contig coverage in both library prep. How were the different quantification results from the two library prep (esp. for the dsDNA viruses that were recovered by both methods) used in the permafrost data to compare across thawing gradient/depth/year? Average? Select one over the other?,

Indeed, we compared the libraries on the contigs and combined the libraries for vOTU generation and downstream analyses. This is stated in lines 324-325 "To explore the ecology of these viruses at the population level, we combined the libraries to create a non-redundant list of viral populations (vOTUs) as reference genomes for further analyses." To make it clearer, we added text in the methods section, lines 158-160 "These dsDNA viral contigs (identified in both library approaches) were clustered at 95% ANI and 85% alignment fraction (AF) to create a final vOTU dataset (Roux et al. 2019b) and their quality was assessed with CheckV default parameters (Nayfach et al. 2020) (Supplementary Table 3)."

16. Line 335, perhaps it corresponds to Fig 3a not 3C? Are we missing some figures or tables to demonstrate the information in lines 337-340?
Our apologies, we had referenced supplementary Table 7 (now Table S9) in the previous sentence that shows the coverage per vOTU and for each sample and did not include it here again. Supplementary Table 9 is now referenced.

17. A clarification question, is the vOTU sharing analysis for both dsDNA and ssDNA viruses or just dsDNA viruses? It may help to make it clear if we specify the types of viruses that vOTUs represent in each analysis (i.e., vOTU sharing analysis, richness/abundance comparison).
We have clarified for each figure which viruses are being analyzed.

18. Fig5 is very eye-catching but a bit hard to read to extract the info that is mentioned in the results and discussion. Maybe we could try to simplify the tree a bit (e.g., collapse some branches)?
We completely understand and appreciate the suggestion. We tried many ways to view this tree but collapsing the tree results in loss of the overall context. To make it easier, we have split the trees into two figures and included them as supplementary data.

19. Could the authors provide a sentence or two to explain how 'Based on our phylogenic comparison of microviruses with those identified in another peatland' can be used to 'infer four microviruses vOTUs to infect bacteria from the phylum Bacteroidota'.
We infer this virus-host interaction due to the close phylogenic relationship between our microviurses and the other peatland microviruses, which are reported to infect Bacteroidota. Additionally, we have updated the text because Bacteroidota are one of the most commonly infected bacterial phyla by microviruses (Kirchberger et al., 2022).

"Based on our phylogenic comparison of microviruses with those identified in another peatland, Les Pradeaux Mire (Fig. 5) (Quaiser et al., 2015), we infer four microviruses vOTUs to infect bacteria from the phylum Bacteroidota. *Microviridae* is a family of ssDNA bacteriophage with a wide host range that mainly infect Bacteroidota (Gram-negative bacteria), *Enterobacteriaceae* (Gram-negative bacteria), *Spiroplasma* (parasitic bacteria without a cell wall), and *Rhizobiaceae* (Gram-negative nitrogen fixing bacteria that have symbiotic relationship with plants and often

dominate *Sphagnum* peatlands), and to a lesser extent other Gram-negative bacteria (Brentlinger et al. 2002; Krupovic and Forterre 2011; Kirchberger et al. 2022; Zucker et al., 2022)."

Kirchberger, P. C., Martinez, Z. A., & Ochman, H. (2022). Organizing the Global Diversity of Microviruses. mBio, e00588-00522.

20. Line 420-422, it seems there is a logic gap between the bloom of ciliate viruses to their roles in carbon cycling. Please add a sentence or two to explain.
Thank you for the suggestion, we have now clarified by providing examples "Ciliates are bacterivorous protozoan that are commonly found in the active layer of permafrost soils, and along with other protozoa play an important role in controlling bacterial assemblages and recycling soil organic matter (Coolen et al. 2011; Schostag et al. 2019). The roles of ciliates (e.g., grazing on bacteria and viruses; DeLong et al., 2023) may increase as permafrost thaw continues and these environments become inundated with water, suggesting their viruses will also become more abundant and play larger roles in carbon cycling (e.g., via the viral shunt)."

DeLong, J.P., Van Etten, J.L., Al-Ameeli, Z., Agarkova, I.V. and Dunigan, D.D., 2023. The consumption of viruses returns energy to food chains. Proceedings of the National Academy of Sciences, 120(1), p.e2215000120.

21. Line 425, which reference is the 'Wu et al. 2022'? there is only one 'Wu et al. 2022' in the reference list that is '108. Wu, R., Smith, C. A., Buchko, G. W., Blaby, I. K., Paez-Espino, D., Kyrpides, N. C., Yoshikuni, Y., McDermott, J. E., Hofmockel, K. S., Cort, J. R. & Jansson, J.K. (2022). Structural characterization of a soil viral auxiliary metabolic gene product–a functional chitosanase. Nature Communications, 13(1), 1-14.'. But it is not about RNA viruses infecting Amoeba. Please check your reference list carefully.

Thank you for the careful observation, indeed, we had the wrong Wu 2022 reference. We have now replaced the old Wu 2022 reference with "Wu, R., Bottos, E. M., Danna, V. G., Stegen, J. C., Jansson, J. K., & Davison, M. R. (2022). RNA viruses linked to eukaryotic hosts in thawed permafrost. Msystems, 7(6), e00582-22." We have reviewed each reference to make sure that it is correct.

22. Did you find any pattern in host compositions across thawing gradient/depth profile/years?
We agree that microbial hosts are an important part of understanding the ecosystems and are the focus of several papers from the consortium. Therefore, we did not look at their patterns here except to characterize inferred virus-host interactions by leveraging the microbial analyses from Woodcroft et al., (2018).

We did look for patterns in virus-host linkages and have included the following new text "Further analyses revealed an increase in the number of viruses linked to hosts with permafrost thaw and a decrease with depth (Supplementary Table 8)."

23. Line 433, AMGs that 'were previously reported' refer to the ones that have been reported in permafrost or generally in soil? AMGs that 'were previously experimentally verified' are mostly about photosynthesis?

This terminology is what is provided by the DRAM-v analyses. The AMGs detected in our dataset are compared to the DRAM-v database and could be assigned a "k" flag (meaning known) when the gene has been annotated with a database identifier representing a function from a previously identified AMG in the literature. The "experimentally verified" flag (E) is similar to the (K) flag, but the AMG has to be an experimentally verified AMG in a previous study, meaning it has been shown in a host to provide a specific function. Many experimentally verified AMGs are related to photosynthesis, but no photosynthesis AMGs were detected in our dataset. Most of the AMGs with an "E" flag in our dataset are glycoside hydrolases.

We added the following text to the manuscript for clarity "If the AMG had been previously identified in the literature according to the DRAM-v database, it received a "K" flag (known) and if the AMG was known and had also been experimentally verified in a previous study (meaning it has been shown in a host to provide a specific function), it received an "E" flag and was labeled "experimentally verified". The remaining AMGs were labeled as "putative AMGs"."

24. Fig 7D: a nice summary of the GHs.

Thank you.

Line-by-line

109: Optimized methods – citations need to be updated to the methods used by the authors

We removed the citations from this section since they are referenced in the previous paragraph and citated elsewhere in the document.

151: typo, citation is in the word, then citation is not upper cased

Done.

298: underlaid?

Underlain is the correct past tense for underlie. Underlaid is the past for underlay, which is not what is meant.

314: is amplified the right word here? To me it implies introducing bias

We feel amplify is the correct word because it is commonly used to mean more DNA.

320: which ones are the PCR amplified metagenomes? Not specifically stated in text

We changed “the” to “these” since all were PCR amplified.

333: which ones are tagmentation based? See above

We added to the text (i.e., all Nextera amplified genomes).

631: interesting – any speculation on what the use of such a gene would be to viruses?

We think the gene allows more control of hosts lysis. We have added this speculation as “We speculate these viruses carry RacD AMGs to ensure a sufficient supply of D-aspartate, an important building block for the maintenance of the bacterial peptidoglycan cell wall. By controlling the availability of D-amino acids, the virus prevents premature lysis of the host cell, which would release incomplete or defunct virions (Kannoly et al., 2022; Guliy and Evstigneeva 2025). This control allows the virus to maximize the duration of its productive cycle, ensuring an optimal and functional virion yield.”