

Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape

Corresponding Author: Professor Corina Brussaard

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Piedade's "Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape," underscores the vulnerability of the Southern Ocean microbial ecosystem to climate change, revealing a previously unknown viral landscape and highlighting the significance of Nucleocytoviricota viruses in regulating phytoplankton dynamics, while also showing complex seasonal patterns in viral populations. While the paper is very descriptive in nature, it provides much needed fine-grade temporal analyses of viral dynamics in a key oceanic region. Overall, the paper is well-written and I only have minor comments below:

Lines 66-73: Please include the dereplicated # of vOTUs here. It helps the reader know the true scope of diversity from the get-go. I know that 0.039% clustered down further in to vOTUs according to methods, but it's important to help compare the paper to other papers.

Lines 161-168: There is not enough introduction to this section. What portion of the vOTUs that you are looking at are putatively temperate? Can you use other tools such as DeePhage further to confirm that they're temperate?

Lines 184-208: The binning performed for this section has high levels of cellular contamination. Given the importance of the novelty of these viruses for this paper, it would be useful for understanding the true length of some these genomes. I suggest using Phables or some other tool to help bridge the gap between genomic fragments to remove some contamination.

Reviewer #2

(Remarks to the Author)

This study concerns a survey of virus communities in the Southern Ocean, using metagenomics and phylogenetics approaches. The main goal is to survey the diversity of viruses in this region, highlighting that these microbial communities remain vastly understudied.

The paper is nicely written, with clear motivation for the work and largely justifiable conclusions. Although the study mainly sets the stage for future work in this compelling system, it provides a valuable survey of the Southern Ocean 'virome'. Below I provide some minor concerns that the authors might address to improve their study, including important clarifications.

Minor concerns

-Some of the language in the paper is imprecise in my opinion. For example, the abstract states that Nucleocytoviricota "regulate" phytoplankton population dynamics. Whereas parasite/predator regulation of host/prey populations is well known in ecology, I am unclear how this is shown in the current study. That is, it seems far more accurate to state that these viruses seem to affect their host populations, but more work should be done when claiming regulation per se. I do not think the current study demonstrates this ecological phenomenon. Also in the abstract, it is stated that an "interplay" between phages and eukaryotic hosts is shown in the study. Just because changes in these microbes are observed to fluctuate over time I do not see any evidence that they are directly influencing one another, as suggested by "interplay".

-Results, Line 102. It would be useful to briefly state why the study's library preparation methods bias against ssDNA viruses, aside from simply citing a reference here.

-Results, Lines 105-106. I understand the claim that ssDNA viruses are likely less common than dsDNA ones, and I do not dispute this is probably true. But the authors are generally claiming that the survey is complete, and this cannot be true if they made no attempt to examine RNA viruses. Again, there could be a claim that these too are relatively rare in the study

environment. But unless I missed it, the authors did not openly state that RNA viruses are necessarily missing here because they did not look for them.

-Results, Line 129. I believe 'Mantel' should be capitalized.

-Results, Lines 220-222. This is a compelling result that suggests the timing of sampling affects the likelihood of virus discovery, owing to seasonality differences in virome composition. That is, if I understand this result correctly, Assuming yes, the authors might want to emphasize this point as it relates to the value of longitudinal sampling across the year, as opposed to sampling at a single point in time, if virus discovery in this system is an explicit goal.

-Discussion, Lines 258-259. Perhaps I am confused because I am not overly familiar with the cited reference #52. But why is this observed NCVs and prasinovirus co-occurrence necessarily parasitism? This seems to assume that the earlier observed association defines the nature of symbiosis in the current study, without explicitly testing for it. Seems more accurate to simply state that the viruses are co-occurring, without claiming parasitism as opposed to mutualism (or no interaction at all).

-Discussion, Lines 277-278. Echoing my above concern with language in the abstract, here the claim again is that regulation has been demonstrated. Maybe I am misunderstanding, and the authors need to clarify. But how is this shown, whereas use of a less-specific phrase should suffice, such as "...role in affecting..."?

Reviewer #3

(Remarks to the Author)

Piedade et al. present a comprehensive metagenomic analysis of the viral community in the Southern Ocean, mainly focusing on dsDNA viruses, over the productive seasons covering almost a one-year timeline. Authors conduct detailed analyses that target phages, nucleocytoviruses (NCVs), polinton-like viruses (PLVs), and virophages, independently. The study identifies several novel species and higher taxonomic clades, notably unveiling a highly diversified PLV. Temporal samples improve the understanding the dynamics of various viruses in the Southern Ocean, for NCVs, PLVs and phages. A targeted search also indicates a domination of temperate phages during the bacterial bloom season. Overall, this work provides an comprehensive dataset that addresses the scarcity of viromic data in the Southern Ocean. The figures are compelling. However, I do have some serious concerns, specifically regarding some of the bioinformatic analyses and interpretation. And the current sample resolution makes it difficult to support the conclusion that NCVs regulate phytoplankton population dynamics. These concerns should be addressed before the manuscript is considered for publication.

Major ones:

1. One of the most significant claims made in this study is that NCVs regulate the phytoplankton population. This conclusion is drawn from observations of the decline of *P. antarctica* and the increase in mesomimiviruses in December 2019. My concerns are as follows: 1) Two samples are insufficient for a claim of "regulation" in terms of resolution; 2) A decline in Prymnesiophyceae (it is unclear whether all Prymnesiophyceae in Fig. S2C are *P. antarctica*) was also observed in February 2019, without correlated dynamics in mesomimiviruses, suggesting that other factors may be "regulating" the population of Prymnesiophyceae.
2. Similarly, from the present seasonal patterns, understanding the interplay between phages, eukaryotic viruses, and potential hosts remains challenging. Mainly because 1) Environmental variables, compared to the interplay with viruses, might have relatively profound influence on microbial seasonal patterns, should be taken more seriously in analyses; 2) Conclusions about interplay lack quantitative support, statistical test, but mostly rely on plain descriptions. I would recommend add more analyses and statistical test. For current methods used to infer relationship between microbial and viral communities, such as the Mantel test, should be more thoroughly described and justified
3. Another highlight of this study is the identification of many novel viral clades, mainly through phylogenetic analysis. However, trees in the manuscript lack the support values (bootstrap values, as mentioned in the methods), and some novel clades, especially crassphage, exhibit long branches. Therefore, I recommend 1) incorporating the support values into the visualizations to demonstrate the confidence in the branching; 2) adding additional evidence, such as gene sharing networks, in a higher resolution, to further confirm whether the genomes in a given new clade cluster together and separate from other clades.
4. Related to the comment#3, when the authors introduce the novel Antarctic lineage of crassviruses, they do not incorporate other marine environmental data from previous surveys, but relying on 245 reference genomes/isolates from ICTV and 673 human-realted genomes/genes. This omission makes it hard for me to determine whether these new clades are truly novel and specific to Antarctica. Therefore, I recommend including bigger environmental dataset (e.g. IMG/VR) of crassphages in their phylogenetic analysis to provide a more comprehensive and accurate representation of these clades' novelty and geographic distribution.
5. The method used to calculate relative abundance influences data interpretation. In this study, authors use method as "by dividing the read counts by the total number of viral reads in each sample..." (Lines 592-595). This normalization approach might be fine if the total abundance of viruses were homogeneous across samples. However, in this dataset, the sum of viral abundance varies largely (Fig. 2C). And phages dominate the viral abundance. Using the total viral reads of a single sample as the denominator could be problematic, potentially exaggerating the relative abundance of viruses in communities where they are absolutely scarce. I recommend that the authors check the reliability of relative abundance, discuss the potential caveats of normalization method and how it could bias the interpretation of the results.
6. The high detection ratio of Mamonoviridae with NCVs is abnormal. And 11.2 Mb is too large to be viral genomes, even

considering the interpretation of EVE. The largest EVE in the literature is about 1.9 Mb [<https://pubmed.ncbi.nlm.nih.gov/33208937/>]. So, the detection threshold for NCVs (particularly for Mammonoviridae) may be not reliable. The detected Mammonoviridae sequence likely be bona-fide eukaryotic sequences. Defining NCV or gEVE requires more viral genes and other genomic evidence. This problematic identification calls for a more rigorous approach. Therefore, it is recommended that the authors carefully check the detected NCV genomes and refine their approach of detection.

7. LN 81-85: Using relative abundance to estimate the the lytic and lysogenic infection is tricky as I mentioned in #5. The term "different temporal occurrences" requires a clear definition.
8. LN 86-87: The sentence "archaeal viruses, which showed higher abundance in November and March" appears to be inconsistent with the data presented in Fig.1E, which indicates a very low abundance of archaeal viruses in March. "Haloviruses" is the only visible taxonomy in Fig.1E, other two colors represent higher abundance than haloviruses. Could members in "Other caudoviricetes" and "Other" be classified into some major groups? And I would also recommend specify the scientific taxonomic name of archaea group, instead of "Haloviruses".
9. LN 92-94: Reference to Fig.5C (cellular fraction, that same to Fig.1D) reveals that there is no clear spike in NCV abundance in March 2018. This observation might stem from methodological inconsistencies, namely the use of contigs or bins.
10. LN 96: I would appreciate if authors provide checkV values for the completeness of genomes.
11. LN 100-101: I am confused by "high abundance" regarding PLVs. As their abundance does not appear high compared to ratios in cellular size fractions or to other viruses in the viral size fraction. Related to it, the interpretation, active infectious, should also be refine.
12. LN 117 Cluster C shows two patterns of Chl-a concentration within the cluster (Fig.2A).
13. Host niche largely influences the seasonal pattern of viruses. Also, given the bias towards the phage community (e.g. 96% LN 133), I recommend ecological analyses (such as in Fig.2A) to be done by prokaryotic and eukaryotic viruses, sperately. Such seperation could provide more accurate insights into virus-host interactions.
14. LN 129: The methodology for conducting the Mantel test requires clarification. Does "microbial" include all four cellular categories? Given the significant impact of environmental variables, a partial Mantel test would be more appropriate. Maybe authors have done it, however, I didn't find the details of the Mantel test.
15. LN 178-179 Fig.3C should be Fig.4C. Core genes are most conserved genes within related genomes, and the statement that "30 core genes could be detected in at least one scaffold" suggests that the majority of pahge scaffolds are highly fragmented. I would recommend to refine the usage of "core genes" and give information that show how complete of these pahges based on the core gene set.
16. LN 197-198: Related to the general comment#3. Raphidovirus usually has long branch in the phylogeny, the description of 4 Raphidoviruses should be carefully given based on the support values.
17. LN 224: The use of promoter motifs for prediction appears to be complex in taxonomy (i.e., the NCV promoters are scattered). So I am not convinced by the effectiveness of this prediction method. I would appreciate authors give more support between predicted virophage and NCVs, and include the confidence level for each prediction.
18. LN 234: MCP duplication and trplication are indeed interesting. However, the synteny plot of Fig.6C shows most multiple copied MCPs being contiguous. This suggests the possibility that duplication may not be biological but rather a technical artifact introduced by the gene calling program (Prodigal -meta) and some sequence fragments, like introns. This observation raises the need for careful evaluation to discern between true biological replication events and artifacts.
19. LN 562 NCVs were also detected at contig levels based on the context. Are these contigs used for the binning in method section "Nucleocytoviricota viruses" or all contigs used for binning. Please refine the methodology.

Minor concerns:

20. LN 71: The 75% novel viral sequences is higher than the ~30% unique populations in the Antarctic as shown in the GOV paper (<https://pubmed.ncbi.nlm.nih.gov/31031001/>). This is expected because the SO samples were scarce. A species rarefaction curve should be included to tell if the viral richness has reached a plateau.
21. Please define early, mid, and late summer with specific months and mark these periods in the plots for clarity.
22. LN 68: Please define the 35%. What about the ratio of reads could be mapped?
23. LN 147 Fig.1F is not for archaeal viruses. Should be Fig1E.
24. LN 199 Fig.S8 should be Fig.S7, and the legend of Fig.S7 is missing.
25. It seems the information between Fig.S9 (v-Contact2) and Fig.1 are redundant. Same to Fig.2C and Fig.3C (Viral particle counts); Fig.5C and Fig.1D.
26. Please use NCVs instead of NCLDVs in Fig.6C.
27. LN 617 The percentage is not clear. Are multiple copy of TerL, MCP and portal within one contig counted multiple times or once?
28. LN 623 It seems that the scaffolds "out of" all ICTV-approved family clades are novel sequences.
29. It was difficult read the supplementary table 2. Please considering use one spreadsheet to include descriptions of the column headers for all tables.
30. Please provide the taxonomic lineage information for each accession in supplementary table 3.
31. The supplementary table 4 seem to be wrong. There are only 17 reference genomes in it and they don't cover the full diversity of NCVs.
32. I will appreciate authors provide metadata for samples.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed my comments.

Reviewer #2

(Remarks to the Author)

The authors nicely addressed my minor concerns, and I have no further recommendations for revision.

Reviewer #3

(Remarks to the Author)

I'd like to thank authors for the effort in addressing my comments. Most of the responses are adequate and directly address my concerns. The authors have made changes where feasible and provided clear explanations. The improvements, such as the partial Mantel test and adding support values, make the results easier to interpret. The adjustments to the manuscript have made the data analysis more convincing. Overall, the manuscript is much improved. Good luck with the data mining on crassphages in the large database.

I still have few comments on the NCV and comtamonation.

1. Concerning comment #3 by reviewer #1, about using additional tools to define the viral genomes and gaps: I noticed that the authors used ViralRecall to decontaminate the cellular sequences. However, the ViralRecall scores, which have the function to define the viral regions by considering the penalty from cellular signals, were not provided in the main text or Table S2. Including the ViralRecall scores could help define the NCV regions and thus address the concerns.

2. Related to the comment above and my original comment #6: "The high detection ratio of Mamonoviridae with NCVs is abnormal...". I selected the RNAPL (GVOGm0023) sequence of 2-643266.cc.b31 (the longest bin) and did a quick blastP analysis. The top hits in the NCBI RefSeq and NR databases are all from cellular organisms, with identity ranges from 37%-50% (RefSeq) and 49%-59% (NR). When I restricted the search to the Nucleocytoviricota (taxid:2732007) in the NR database, it yielded hits with identity ranges from 26%-37%. Although this may be due to incorrect labels in the NCBI databases, and the blastP identity doesn't necessarily represent evolutionary relatedness, I am unable to judge whether those mamonovirus markers are viral, or they are originally cellular genes. We know that viral RNAPs resemble eukaryotic RNAPs (I and II) [DOI: 10.1073/pnas.1912006116]. Therefore, an HMM-based detection may yield false positives on some eukaryotic marker genes, leading to false positives in giant virus bin detection, especially given that authors used two marker genes as the cut-off for this clade.

3. Some taxa should be in italic: LN244 "Mamonoviridae," LN245 "Mirusviricota," the legend in Fig 5A, and LN909 "Nucleocytoviricota" and "Preplasmiviricota."

4. LN235. Should "late and early season" be changed to "...early summer..."? I see that there are six genomes represent an "early summer only" seasonality pattern, but I don't understand what "late" stands for.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

I would like to thank the authors for addressing my comments. I'm pleased that my suggestions regarding the identification of novel virus-like sequences (Mriyaviricetes) were helpful. The manuscript has been much improved, and I have no further concerns.

One small detail: it seems that a reference may be duplicated (#101 and #108).

Open Access This Peer Review File 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 license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <https://creativecommons.org/licenses/by/4.0/>

Thank you for allowing us to improve our manuscript. We are grateful for the Reviewers comments, which we have taken into consideration to improve the manuscript. We were also pleased to see that the reviewers shared our excitement about the findings of this study.

Below we answered each of the comments carefully (our replies in italic font). We feel that the edits have significantly strengthened and improved our story. Specifically, we have:

- Edited and clarified the text, removing unwarranted statements about cause/consequence, such as the NCVs “regulating” plankton communities. The language has been adjusted to more accurately reflect our findings.
- We have clarified the justification and the consequences of the used normalizations.
- We have improved the statistical analysis, implemented the partial Mantel test and also sub setting these into phage-prokaryotes and NCVs-eukaryotes.
- Added the support values to the trees of all main figures and a more careful taxonomic classification for NCVs. Included a virophage and PLV gene content analysis in a new Supplementary Figure 8.
- Performed several new analyses (such as testing multiple phage lifestyle tools, mining the IMG/VR for crassphage TerL genes, and PLV and virophage gene content clustering, both for the Rebuttal and manuscript).

We believe that the revised manuscript has been significantly strengthened and look forward to your response.

Yours sincerely,

Goncalo Piedade and Prof. Dr. Corina Brussaard (on behalf of the coauthors)

Reviewer #1 (Remarks to the Author):

Piedade’s “Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape,” underscores the vulnerability of the Southern Ocean microbial ecosystem to climate change, revealing a previously unknown viral landscape and highlighting the significance of Nucleocytoviricota viruses in regulating phytoplankton dynamics, while also showing complex seasonal patterns in viral populations. While the paper is very descriptive in nature, it provides much needed fine-grade temporal analyses of viral dynamics in a key oceanic region. Overall, the paper is well-written and I only have minor comments below:

1 -Lines 66-73: Please include the dereplicated # of vOTUs here. It helps the reader know the true scope of diversity from the get-go. I know that 0.039% clustered down

further in to vOTUs according to methods, but it's important to help compare the paper to other papers.

We have now added that information to the results: “The mid- to high-quality viral sequences clustered into 7942 vOTUs” (line 75).

2 -Lines 161-168: There is not enough introduction to this section. What portion of the vOTUs that you are looking at are putatively temperate? Can you use other tools such as DeePhage further to confirm that they're temperate?

Our approach was to look for the presence of genes known to be involved in lysogeny. We have added a short introduction to this section “High prevalence of lysogenic infection has been reported throughout the Southern Ocean, and are potentially involved in the overwintering survival of Antarctic bacteriophages” (lines 172-173) and the amount of vOTUs that have markers of lysogeny “Of the mid to high-quality phage sequences, 386 had at least one of the lysogeny genes.” (line 177). Following the reviewer’s suggestion, we have tested a few tools such as DeePhage, PhaTYP and Bacphlip. Both DeePhage and PhaTYP rely on deep learning and aim to give a prediction based on small fragments. Bacphlip is a Random Forest classifier based on conserved protein domains of lysogeny genes. We found that DeePhage and PhaTYP performed rather poorly in the identification compared to the presence of markers of lysogeny, in mid-high quality viral sequences (Table R1). Bacphlip predicted similar phages to the paper’s approach. The temporal dynamics are also highly similar (Fig. R1), so we have not added these results to our manuscript. Of course, if the Reviewer insists, we could additionally report the Bacphlip predictions, but we think it is redundant.

Table R1 – Phage lifestyle prediction for mid to high-quality phage sequences (n= 7527)

		Lysogeny markers (this paper)	PhaTYP	DeePhage	Bacphlip
Temperate	Agrees with marker presence	386	185	257	314
	Virulent w/ lysogeny marker		198	129	72
Virulent	Agrees with no absence	-*	5740	4363	7096
	Temperate w/o lysogeny marker		1181	2778	45
No prediction		7141*	190		

* Absence of proof is not proof of absence

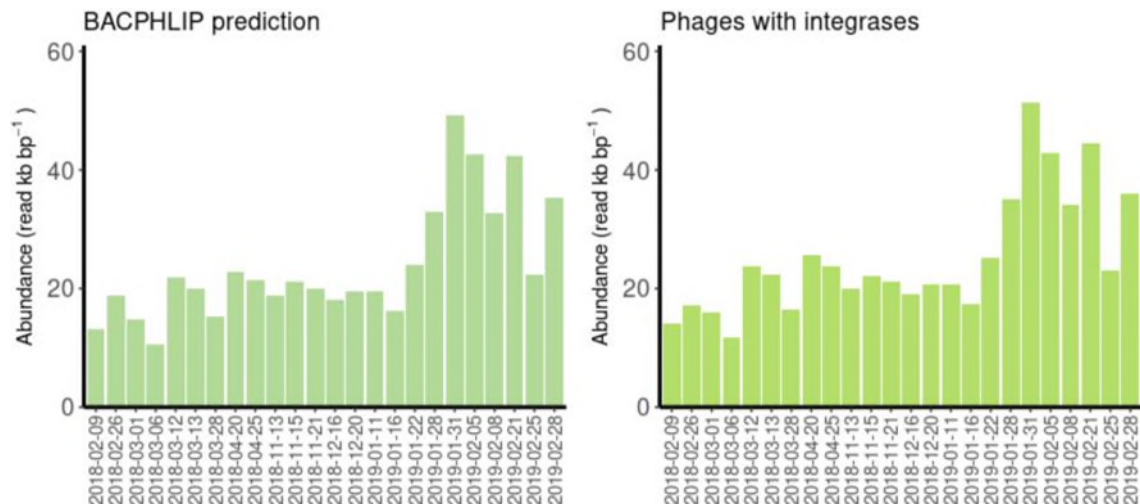


Figure R1 – Comparison of the BACPHILIP lifestyle predictions (left) with the paper gene content-based predictions from manuscript Fig. 3e (right).

3 -Lines 184-208: The binning performed for this section has high levels of cellular contamination. Given the importance of the novelty of these viruses for this paper, it would be useful for understanding the true length of some these genomes. I suggest using Phables or some other tool to help bridge the gap between genomic fragments to remove some contamination.

We attempted to further separate the NCV signal, but the fragments were too short and many with both Eukaryotic and NCV signal (Supplementary Table 2). The tools the reviewer suggests are designed specifically for phages and not suitable for NCVs and, as far as we know, there are no equivalent tools compatible or tested for NCVs. Some NCVs have been found to integrate into their host genomes [1,2], and that might be the reason why we find some chimeric Eukaryotic/NCV bins. We have further discussed and clarified our interpretation “Fourteen of these bins also had high eukaryotic signal, and 10 contained scaffolds with both eukaryotic signal and NCV phylogenetic markers (Supplementary Table 2). The co-occurrence of these signals and the unusually large size of the bins suggest these putative NCVs could be inserted in the genomes of their host as endogenous viral elements. However, other biologic or technical explanations, such as transposable element mediated gene transfer or chimeric binning, are also possible.” (lines 224-229). For the purposes of our analysis, i.e., focusing on viral phylogenetic marker genes, it is not an issue to have bins that are a mix of Eukaryotic and NCV sequences.

Reviewer #2 (Remarks to the Author):

This study concerns a survey of virus communities in the Southern Ocean, using metagenomics and phylogenetics approaches. The main goal is to survey the diversity of viruses in this region, highlighting that these microbial communities remain vastly understudied.

The paper is nicely written, with clear motivation for the work and largely justifiable conclusions. Although the study mainly sets the stage for future work in this compelling

system, it provides a valuable survey of the Southern Ocean 'virome'. Below I provide some minor concerns that the authors might address to improve their study, including important clarifications.

Minor concerns

1 -Some of the language in the paper is imprecise in my opinion. For example, the abstract states that Nucleocytoviricota “regulate” phytoplankton population dynamics. Whereas parasite/predator regulation of host/prey populations is well known in ecology, I am unclear how this is shown in the current study. That is, it seems far more accurate to state that these viruses seem to affect their host populations, but more work should be done when claiming regulation per se. I do not think the current study demonstrates this ecological phenomenon. Also in the abstract, it is stated that an “interplay” between phages and eukaryotic hosts is shown in the study. Just because changes in these microbes are observed to fluctuate over time I do not see any evidence that they are directly influencing one another, as suggested by “interplay”.

We understand the reviewer’s viewpoint. We have amended the wording to more precisely convey this in various instances throughout the manuscript:

- *“highlighting their potential as important regulators of phytoplankton population dynamics” (line 32)*
- *“which underscores the apparent interactions with their microbial hosts” (lines34-35)*
- *“The prevalence and temporal dynamics of these NCVs reinforces their active role in affecting Antarctic phytoplankton population dynamics (lines 318-319)*
- *“their role in impacting bloom decline of the co-occurring P. antarctica host (reductive control), aligning with recent findings of viral lysis being a major mortality factor of this Antarctic phytoplankter” (lines 324-325)*

2 - Results, Line 102. It would be useful to briefly state why the study’s library preparation methods bias against ssDNA viruses, aside from simply citing a reference here.

We have adapted the text to “Though the used standard Illumina library preparation method is biased against ssDNA, we identified...” to clarify this (line 109).

3 - Results, Lines 105-106. I understand the claim that ssDNA viruses are likely less common than dsDNA ones, and I do not dispute this is probably true. But the authors are generally claiming that the survey is complete, and this cannot be true if they made no attempt to examine RNA viruses. Again, there could be a claim that these too are relatively rare in the study environment. But unless I missed it, the authors did not openly state that RNA viruses are necessarily missing here because they did not look for them.

*In line 112 of the revised manuscript, we now clarify that the presented results on the observed ssDNA virus diversity allows an important view of their diversity, making our survey more complete regarding **DNA** virus diversity. Our study characterizes the DNA virus*

diversity and to further clarify this, we also replaced “viruses” by “DNA viruses” in the Abstract.

4 - Results, Line 129. I believe ‘Mantel’ should be capitalized.

We corrected that accordingly.

5 - Results, Lines 220-222. This is a compelling result that suggests the timing of sampling affects the likelihood of virus discovery, owing to seasonality differences in virome composition. That is, if I understand this result correctly, Assuming yes, the authors might want to emphasize this point as it relates to the value of longitudinal sampling across the year, as opposed to sampling at a single point in time, if virus discovery in this system is an explicit goal.

We appreciate the reviewer’s recommendation and made it an additional point in our discussion: “The comprehensive seasonal coverage and enhanced sampling resolution of this study reveals that timing of sampling affects the likelihood of virus detection and discovery.” (lines 330-331 of the revised manuscript). We also included a similar conclusion by reviewer 3 on the topic.

6 - Discussion, Lines 258-259. Perhaps I am confused because I am not overly familiar with the cited reference #52. But why is this observed NCVs and prasinovirus co-occurrence necessarily parasitism? This seems to assume that the earlier observed association defines the nature of symbiosis in the current study, without explicitly testing for it. Seems more accurate to simply state that the viruses are co-occurring, without claiming parasitism as opposed to mutualism (or no interaction at all).

All virophages isolated and characterized to date depend on a NCV in such a way that could be considered parasitic to their NCV “host” [3] and at least for Cafeteria burkhardae mutualistic to the eukaryotic host [4]. We have changed “parasitism” to “depend on”. We have also clarified that we refer to the promotor motif matching analysis here “In our analysis, we attempted to match virophages to their co-occurring NCVs by their putative promoters. Surprisingly, we found promoter signals that were shared between virophage and prasinovirus sequences, suggesting that virophages can depend on NCVs outside the Imitervirales order.” (lines 296 to 299).

7 - Discussion, Lines 277-278. Echoing my above concern with language in the abstract, here the claim again is that regulation has been demonstrated. Maybe I am misunderstanding, and the authors need to clarify. But how is this shown, whereas use of a less-specific phrase should suffice, such as “...role in affecting...”?

We agree and have changed it accordingly.

Reviewer #3 (Remarks to the Author):

Piedade et al. present a comprehensive metagenomic analysis of the viral community in the Southern Ocean, mainly focusing on dsDNA viruses, over the productive seasons covering almost a one-year timeline. Authors conduct detailed analyses that target phages, nucleocytoviruses (NCVs), polinton-like viruses (PLVs), and virophages, independently. The study identifies several novel species and higher taxonomic clades, notably unveiling a highly diversified PLV. Temporal samples improve the understanding the dynamics of various viruses in the Southern Ocean, for NCVs, PLVs and phages. A targeted search also indicates a domination of temperate phages during the bacterial bloom season. Overall, this work provides an comprehensive dataset that addresses the scarcity of viromic data in the Southern Ocean. The figures are compelling. However, I do have some serious concerns, specifically regarding some of the bioinformatic analyses and interpretation. And the current sample resolution makes it difficult to support the conclusion that NCVs regulate phytoplankton population dynamics. These concerns should be addressed before the manuscript is considered for publication.

Major ones:

1. One of the most significant claims made in this study is that NCVs regulate the phytoplankton population. This conclusion is drawn from observations of the decline of *P. antarctica* and the increase in mesomimiviruses in December 2019. My concerns are as follows: 1) Two samples are insufficient for a claim of "regulation" in terms of resolution; 2) A decline in Prymnesiophyceae (it is unclear whether all Prymnesiophyceae in Fig. S2C are *P. antarctica*) was also observed in February 2019, without correlated dynamics in mesomimiviruses, suggesting that other factors may be "regulating" the population of Prymnesiophyceae.

We have addressed the reviewer's concern and changed "regulating" to "affecting" following the suggestion of reviewer 2 point 7 (lines 319). We now state more carefully the potential role of the Tethysvirus: "Our study's temporal resolution shed light on their role in impacting bloom decline of the co-occurring P. antarctica host (reductive control)" (lines 324-325). Additionally, we have specified the share of P. antarctica in the Results: "coincided with a decline in P. antarctica, the most abundant Prymnesiophytehyceae (>99% of Prymnesiophyceae rRNA gene reads, except on the 9th February and 6th March 2018)" (lines 210-211).

2. Similarly, from the present seasonal patterns, understanding the interplay between phages, eukaryotic viruses, and potential hosts remains challenging. Mainly because 1) Environmental variables, compared to the interplay with viruses, might have relatively profound influence on microbial seasonal patterns, should be taken more seriously in analyses; 2) Conclusions about interplay lack quantitative support, statistic test, but mostly rely on plain descriptions. I would recommend add more analyses and statistical test. For current methods used to infer relationship between microbial and viral communities, such as the Mantel test, should be more thoroughly described and justified

We phrase 'regulation' and 'interplay' now more carefully (see also our replies to reviewer 2 point 1). We employ a multivariate statistical model to accurately quantify the explanatory potential of the measured environmental variables on the composition of viral communities. We used the partial-Mantel test to quantify the correlation between the phage and bacterial

communities. As per the reviewer suggestion in point 14, we now incorporate also a paired-Mantel test that accounts for the effects of environmental variables. We have detailed these in the Methods: “The association between the phage and bacterial communities were tested by using a partial Mantel test on the Aitchison distance matrixes, while controlling for environmental conditions using the *mantel.partial* function (Pearson correlation and 999 permutations, *Vegan* v2.6-4). The control environmental matrix was computed as the Euclidean distances of the non-covarying variables which were significant according to the general multivariate regression model.” (lines 798-802).

3. Another highlight of this study is the identification of many novel viral clades, mainly through phylogenetic analysis. However, trees in the manuscript lack the support values (bootstrap values, as mentioned in the methods), and some novel clades, especially crassphage, exhibit long branches. Therefore, I recommend 1) incorporating the support values into the visualizations to demonstrate the confidence in the branching; 2) adding additional evidence, such as gene sharing networks, in a higher resolution, to further confirm whether the genomes in a given new clade cluster together and separate from other clades.

We have added to all the trees in figures 4,5 and 6 a green star that indicates a node support higher than 85% at the base of the relevant clades. For our phylogenetic analysis we have used well established phylogenetic marker genes. For NCVs these consisted of 7 concatenated markers [5] of which 4 are shared with *Mirusviricota* [6]. For *Crassvirales* we have performed the phylogenetic analysis using the *TerL*, *MCP* and *Portal* proteins that were also used to define taxa in the *Crassvirales* order [7]. For virophages and polinton-like viruses, the major capsid protein is a commonly used phylogenetic or clustering marker [8,9]. These 2 groups share different sets of overlapping genes with diverse evolutionary trajectories [10]. We have produced a gene sharing network for virophages and polinton-like viruses (Figure R2, manuscript Supplementary Fig. 8) which shows that these tend to cluster similarly to the *MCP* phylogeny. We have added the following to the Results: “Their gene-sharing network clustering agrees with the *MCP* clades (Supplementary Fig. 8), while viruses of the *Omnilimnoviridae* seem to share more genes with *TSV* and *PGVV* group *PLVs* than with other virophages. Similarly, *Chi* group *PLVs* seem to share more genes with virophages than with other *PLVs*, reflecting their complex evolutionary relationships.” (lines 240-244). Furthermore, we added “*MCP*” in two instances to the text to clarify that the novelty regards the *MCP* gene: “distinct *MCP* clade” (line 248) and “*MCP* phylogenetic groups” (line 260).

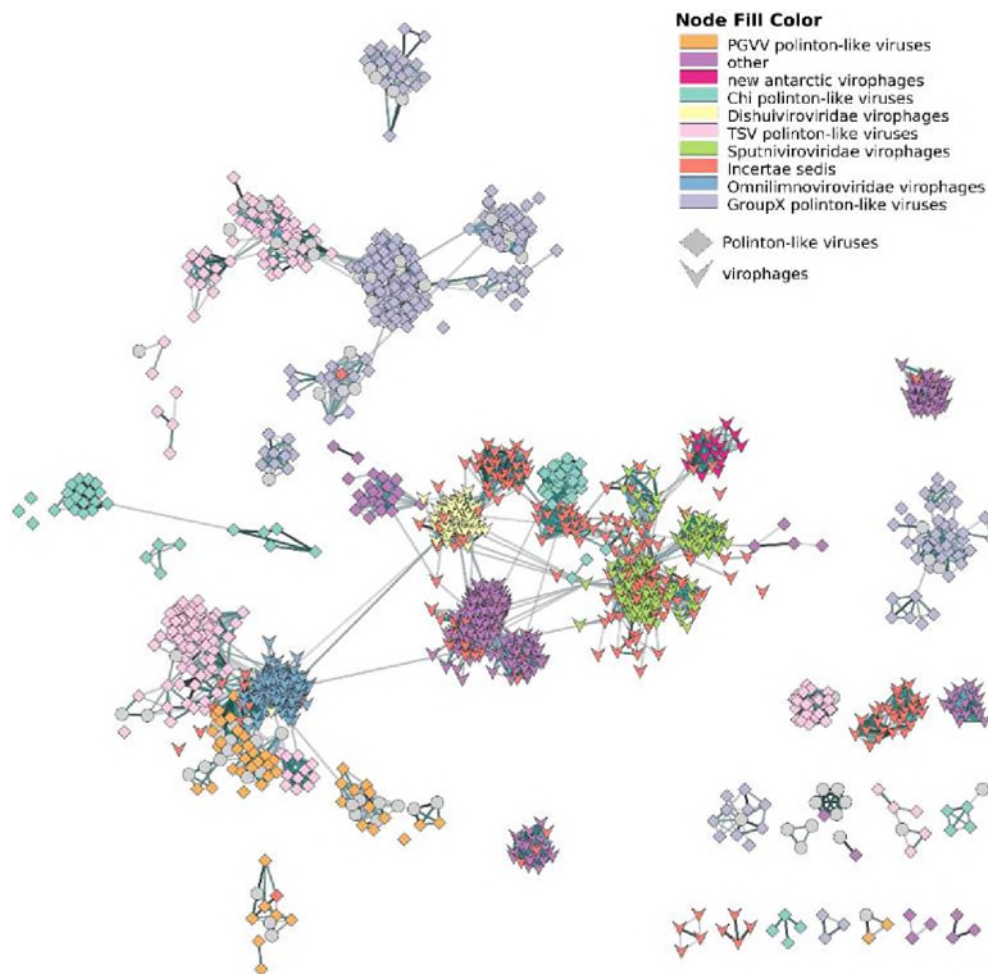


Figure R2 – vContact2 gene sharing network of virophage and PLV sequences longer than 5kb. Diamonds and arrows are respectively PLV and virophage sequences, these are coloured according to the clades defined in figure 6.

4. Related to the comment#3, when the authors introduce the novel Antarctic lineage of crassviruses, they do not incorporate other marine environmental data from previous surveys, but relying on 245 reference genomes/isolates from ICTV and 673 human-related genomes/genes. This omission makes it hard for me to determine whether these new clades are truly novel and specific to Antarctica. Therefore, I recommend including bigger environmental dataset (e.g. IMG/VR) of crassphages in their phylogenetic analysis to provide a more comprehensive and accurate representation of these clades' novelty and geographic distribution.

We agree with the reviewer that the current study cannot tell if these are exclusive to the Antarctic or not. Some of the co-authors are currently mining the IMG/VR for crassphages and writing another manuscript on a broader analysis of this order. They agreed to share some preliminary results which show that many of the clades found in this study are populated by other IMG/VR sequences (Fig R3), suggesting that the lineages that we found are not unique to Antarctica, although they are novel. We think that adding these results to the current publication would be too preliminary and outside the scope of our paper's objectives on the characterization of Antarctic viral diversity. We did add to the Discussion that these results invite further investigation of global non-host associated crassphages: "The new-found diversity of crassphages in Antarctic calls for further investigation of this

class regarding diversity and activity in both the marine and other non-host associated environments. Mining public data repositories may reveal additional members of these newly discovered lineages.” (lines 287-290).

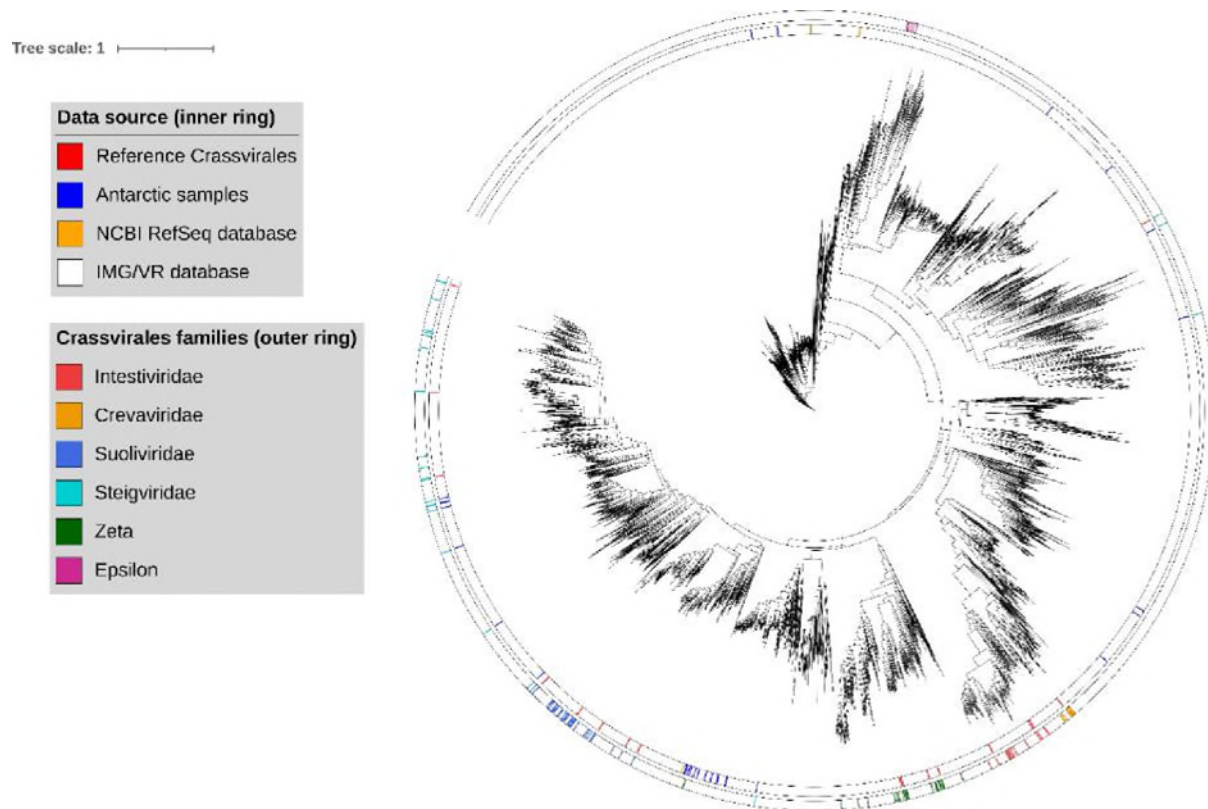


Figure R3 – Phylogenetic tree of the terminase large subunit (TerL) of Crassvirales sequences found in this study (Antarctic samples), references (Yutin et al 2018 and 2021), NCBI RefSeq and IMG/VR. We can see that all clades with Antarctic sequences (blue inner ring) also contain IMG/VR sequences (white inner ring).

5. The method used to calculate relative abundance influences data interpretation. In this study, authors use method as "by dividing the read counts by the total number of viral reads in each sample..." (Lines 592-595). This normalization approach might be fine if the total abundance of viruses were homogeneous across samples. However, in this dataset, the sum of viral abundance varies largely (Fig. 2C). And phages dominate the viral abundance. Using the total viral reads of a single sample as the denominator could be problematic, potentially exaggerating the relative abundance of viruses in communities where they are absolutely scarce. I recommend that the authors check the reliability of relative abundance, discuss the potential caveats of normalization method and how it could bias the interpretation of the results.

Indeed, shotgun metagenomics cannot accurately recover viral abundance, and the data are at best compositional. We have added a short explanation to the figure 1 legend, to help guide the reader in the interpretation: "Note that these are compositional values and to

reflect absolute abundances, relate to the total virus particle abundances (Fig. 3c) ". As for the choice of denominator, although non-viral and unclassified DNA was detected, our sampling methods specifically targeted viruses. Thus, we decided that it would best reflect virus diversity using virus reads as denominator (to remove the effect of non-viral contamination). Despite the best efforts to produce the cleanest viromes, there is always some cellular contamination [12]. We have disclosed the fraction of reads that map to viral contigs, non-viral contigs or to none in Figure S1. To clarify the motivation, we have added to the Methods section: "To remove the effect of non-viral contamination, the read relative abundances were calculated by dividing the read counts by the total number of viral reads in each sample" (lines 655-656). Additionally, we have added the motivation for using the compositional data analysis methods: "Given the compositional nature of the abundance data, the analysis was performed using compositional data analysis methods." (lines 780-781).

6. The high detection ratio of Mamonoviridae with NCVs is abnormal. And 11.2 Mb is too large to be viral genomes, even considering the interpretation of EVE. The largest EVE in the literature is about 1.9 Mb <https://pubmed.ncbi.nlm.nih.gov/33208937/>. So, the detection threshold for NCVs (particularly for Mamonoviridae) may be not reliable. The detected Mamonoviridae sequence likely be bona-fide eukaryotic sequences. Defining NCV or gEVE requires more viral genes and other genomic evidence. This problematic identification calls for a more rigorous approach. Therefore, it is recommended that the authors carefully check the detected NCV genomes and refine their approach of detection.

We understand how our phrasing may have been confusing. We improved the explanation of these results: "Fourteen of these bins also had high eukaryotic signal, and 10 contained scaffolds with both eukaryotic signal and NCV phylogenetic markers (Supplementary Table 2). The co-occurrence of these signals and the unusually large size of the bins suggest these putative NCVs could be inserted in the genomes of their host as endogenous viral elements. However, other biologic or technical explanations, such as transposable element mediated gene transfer or chimeric binning, are also possible." (lines 224-229). We attempted to further separate the NCV signal, but the fragments containing phylogenetic markers were too short and many with both Eukaryotic and NCV signal (supplementary table 2). For the purposes of our analysis, i.e., focusing on viral phylogenetic marker genes, it is not an issue to have bins that are a mix of Eukaryotic and NCV sequences.

7. LN 81-85: Using relative abundance to estimate the the lytic and lysogenic infection is tricky as I mentioned in #5. The term "different temporal occurrences" requires a clear definition.

It was not our intention to estimate lytic and lysogenic infection. We wish to guide the reader on the different possible origins for viral reads in the cellular and viral fraction and how these can lead to the differences in patterns. Given the temporal resolution and the fact that the cellular and viral fraction are from the same sample, we can hypothesize the reasons why we find a virus to be more abundant in the cellular fraction versus the viral fraction and when it increases in both fractions. Ultimately, we cannot make definite conclusions. We have re-written this section for clarity: "Differing abundance patterns in the viral and cellular fraction

can be the result of seasonal dynamics in lytic and lysogenic viral infection, as the viral reads from the cellular fraction can reflect both actively replicating and lysogenic viral infection while the virus fraction is composed of free viral particles.” (lines 83-86).

8. LN 86-87: The sentence "archaeal viruses, which showed higher abundance in November and March" appears to be inconsistent with the data presented in Fig.1E, which indicates a very low abundance of archaeal viruses in March. "Haloviruses" is the only visible taxonomy in Fig.1E, other two colors represent higher abundance than haloviruses. Could members in "Other caudoviricetes" and "Other" be classified into some major groups? And I would also recommend specify the scientific taxonomic name of archaea group, instead of "Haloviruses".

The reviewer is correct (we apologize for the mistake) and we have changed "November and March" to "November to January". (line 92). Furthermore, we have changed the taxonomic name to the order "Thumleimavirales" which more accurately and conservatively reflects the "halovirus" taxonomic name attributed by cenote-taker2. We have added this to Methods: "Archaeal group sequences classified as halovirus by cenote-taker were classified as Thumleimavirales." (line 631-632). The other caudoviricetes in the archaeal vContact2 cluster were unfortunately not classified below class level. We have added a new Supplementary Table 1 with the full information about all the 8045 mid to high quality viral scaffolds.

9. LN 92-94: Reference to Fig.5C (cellular fraction, that same to Fig.1D) reveals that there is no clear spike in NCV abundance in March 2018. This observation might stem from methodological inconsistencies, namely the use of contigs or bins.

The difference between the 2 figures originates not from the use of contigs versus bins but from the use of viral reads versus vertical coverage. Figure 1 uses the number of viral reads in each taxon divided by the total number of viral reads. Figure 5 uses number of viral reads mapping to the NCV bin normalized by bin length and corrected for sequencing depth (total reads). We have added this information to the figure 5 legend: "metagenomes in bin vertical coverage (read abundance normalised by bin length". Additionally, we added to the Methods our reasons to use read abundance versus vertical depth for different analysis: ". Read abundance was used where assembly genome fragmentation could impact abundance such as for higher taxonomy level compositional abundance. Vertical coverage was used where genome fragmentation was not expected to be an issue, such as for scaffold abundance, cumulative abundance of scaffolds containing a single gene, or the binned NCV genomes." (lines 659-663).

10. LN 96: I would appreciate if authors provide checkV values for the completeness of genomes.

Please note that the limited reference database for polinton-like viruses might mean that CheckV is not fine-tuned to accurately estimate the completeness of this group. We have nonetheless added the requested information, including the number of sequences with terminal repeats: "accounting for 1,678 scaffolds, 53 had a CheckV completeness higher

than 50%, and six contained terminal repeats. In total 243 exceeded 10 kb, which likely represent near-complete genomes, as their length is typically around 20 kb.” (lines 101-103)

11. LN 100-101: I am confused by "high abundance" regarding PLVs. As their abundance does not appear high compared to ratios in cellular size fractions or to other viruses in the viral size fraction. Related to it, the interpretation, active infectious, should also be refined.

We have rephrased to “Meanwhile, their abundance and detection in the viral fraction likely indicates a shift to an active infectious state” (line 107-108).

12. LN 117 Cluster C shows two patterns of Chl-a concentration within the cluster (Fig.2A).

We have changed the wording to better reflect the change observed in cluster C: “Late-summer (Cluster C) showed reducing Chl-a concentrations and prokaryote abundance and the peak in viral abundance” (lines 125-126).

13. Host niche largely influences the seasonal pattern of viruses. Also, given the bias towards the phage community (e.g. 96% LN 133), I recommend ecological analyses (such as in Fig.2A) to be done by prokaryotic and eukaryotic viruses, separately. Such separation could provide more accurate insights into virus-host interactions.

The reviewer has a point on the fact that the statistical analysis will mostly recover the prokaryotic signal. We have now split the analysis and for the prokaryotic viruses we rephrased to: “for the prokaryotic community composition and phage community alone the partial Mantel r-statistic was 0.3411 with a p-value of 0.002.” (lines 136-137).

Doing an analysis focusing on the eukaryotic viruses is more challenging than simply separating them. We have performed a similar partial Mantel test for the NCVs and the Eukaryotic fraction: “The cellular fraction NCV composition co-varied with the Eukaryotic composition (partial Mantel r-statistic 0.2888, p-value 0.001), while controlling for environmental variables.” (lines 232-234).

14. LN 129: The methodology for conducting the Mantel test requires clarification. Does “microbial” include all four cellular categories? Given the significant impact of environmental variables, a partial Mantel test would be more appropriate. Maybe authors have done it, however, I didn’t find the details of the Mantel test.

We have performed a partial Mantel test and rephrased the text to: “The overall Microbial community composition explained 68% of the variation in the viral community when controlling for the environmental variables (partial Mantel r-statistic=0.6842, p-value=0.01).” lines (137-139). Accordingly, we have added information to the Methods: “The association between the phage and bacterial communities were tested by using a partial Mantel test on the Aitchison distance matrixes, while controlling for environmental conditions using the mantel.partial function (Pearson correlation and 999 permutations, Vegan v2.6-4). The control environmental matrix was computed as the Euclidean distances of the non-covarying

variables which were significant according to the general multivariate regression model.” (lines 798-802).

15. LN 178-179 Fig.3C should be Fig.4C. Core genes are most conserved genes within related genomes, and the statement that “30 core genes could be detected in at least one scaffold” suggests that the majority of pahge scaffolds are highly fragmented. I would recommend to refine the usage of “core genes” and give information that show how complete of these pahges based on the core gene set.

We have added information regarding the completeness according to CheckV: “Of these, 17 were longer than 50 kb, 18 were at least 50% complete according to CheckV and seven had a completeness higher than 95%, two with direct terminal repeats” (lines 180-182)

16. LN 197-198: Related to the general comment#3. Raphidovirus usually has long branch in the phylogeny, the description of 4 Raphidoviruses should be carefully given based on the support values.

We have taken a more careful approach and renamed the viruses in the Aylward et al. (2021) family level clade AG_04 [5] to “Raphidovirus-like” (line 219). Similarly, those in AG_01 that do not fall strictly inside the Prasinovirus clade were reclassified “candidate family Prasinoviridae AG_01” (line 217). Below we show a more detailed Algavirales tree which has good support values (Fig. R5).

→ Other NCVs

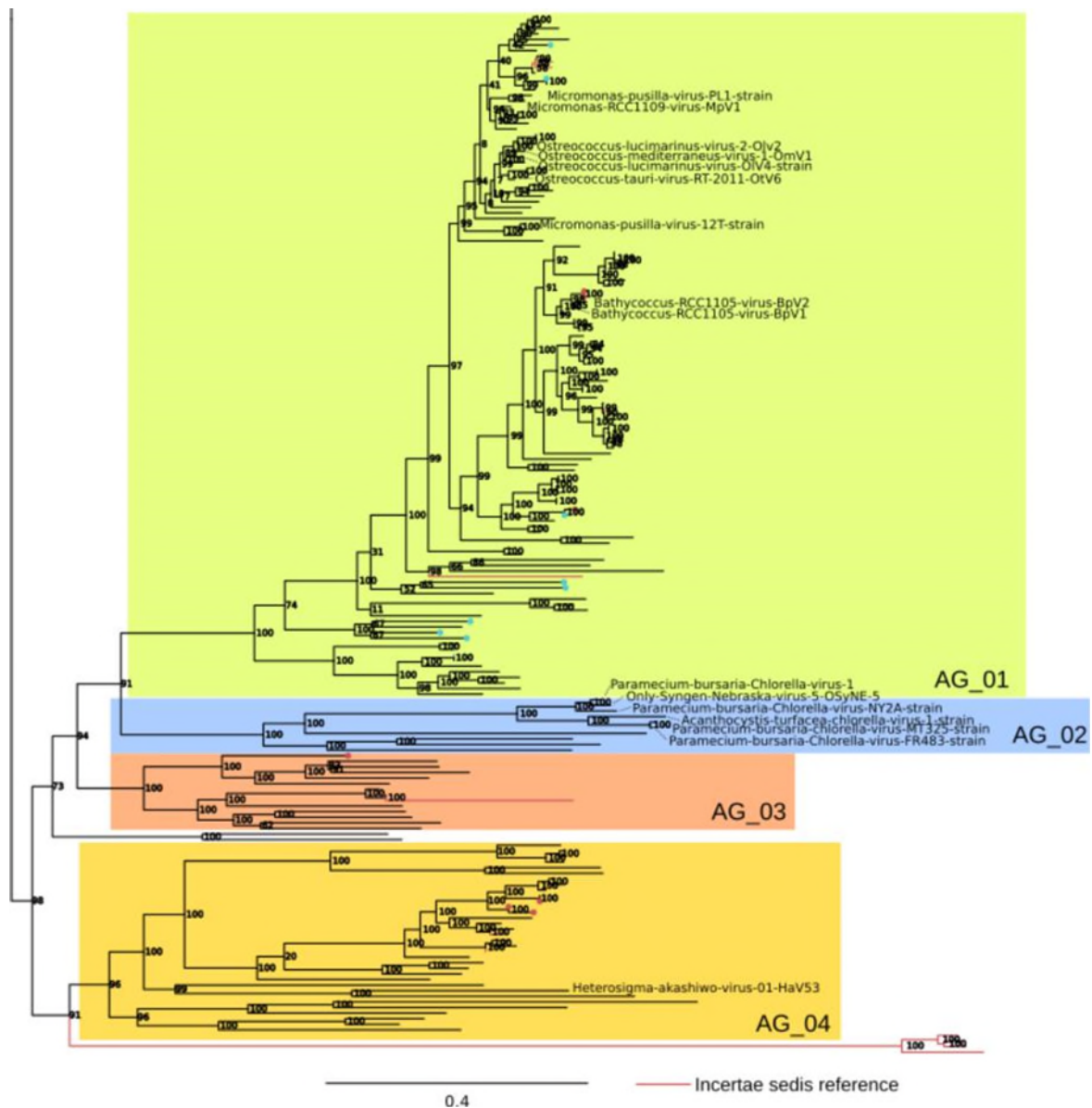


Figure R5 – Phylogenetic tree of the Algavirales order. The clades are colored as per Aylward et al. 2021. Antarctic NCVs from this study are highlighted as circles, red for those detected in the cellular fraction and blue for those detected in the viral fraction. Isolates are indicated in the tree. Bootstrap support values are indicated at the nodes.

17. LN 224: The use of promoter motifs for prediction appears to be complex in taxonomy (i.e., the NCV promoters are scattered). So I am not convinced by the effectiveness of this prediction method. I would appreciate authors give more support between predicted virophage and NCVs, and include the confidence level for each prediction.

The use of the late-promoter motifs can provide us with an idea of who the host might be, similar to the approach by Roux et al. (2017) [11]. We have now added the full results with e-value and true and false discovery rates as the Supplementary Table 4. To further validate our approach, we have added the virophage mavirus and the CroV to the analysis and the

CroV motif comes up as the best match to *mavirus* (Supplementary Table 4). We added to the Discussion a remark that highlights the fact that we do not know how wide-spread the promoter match is: “Virophages replicate with co-infecting NCVs and may, in some cases such as *mavirus*, provide their hosts population-level protection against NCVs^{61,62}. Since virophages parasitize the transcription machinery of their associated NCVs, the gene promoter motifs between giant viruses and virophages often share detectable similarity. In our analysis, we attempted to match virophages to their co-occurring NCVs by their putative promoters. Surprisingly, we found promoter signals that were shared between virophage and prasinovirus sequences, suggesting that virophages can depend on NCVs outside the *Imitervirales* order” (lines 293-299). We also changed in the Results “We identified putative NCVs associated” to “We identified NCVs putatively associated” (line 253), so it is clearer that the association is the prediction.

18. LN 234: MCP duplication and triplication are indeed interesting. However, the synteny plot of Fig.6C shows most multiple copied MCPs being contiguous. This suggests the possibility that duplication may not be biological but rather a technical artifact introduced by the gene calling program (Prodigal -meta) and some sequence fragments, like introns. This observation raises the need for careful evaluation to discern between true biological replication events and artifacts.

Although we understand the point the reviewer is making, we think the MCP duplication is not a technical artifact. If that were the case, these MCP would be 1/3 the size of those from PLVs containing 1 MCP (e.g. the TSV NODE_3408 versus the TSV PLV-SPO1). That is not the case as can be seen in Fig. 6d, indeed suggesting multiple gene copies. Furthermore, visual inspection of the protein alignment shows that these are homologous over the full length (no large gaps). The alignment is available in the data repository associated with the publication.

We also agree with the reviewer that these results are interesting. We found that while in group X the MCP are quite divergent, in group TSV these are usually closely related. Overall, these results suggest these to be true MCP duplications and not technical artifacts. We have expanded in the Results on the MCP triplication differences between group X and TSV: “MCPs from the same PLV are generally more closely related in TSV than in group X, with their phylogenetic distance being on average 2.4 times greater than in group X (p -value < $2.2E-16$, t -test). Additionally, this distance is on average 2.26 times smaller than between MCPs from different PLVs in the TSV group (p -value < $2.2E-16$, t -test).” (lines 265 to 268). And we added to the Discussion: “The high divergence between the group X MCPs suggests this to be an early evolutionary event, contrasting with the triplication in group TSV which tends to have more closely related triplicates.” (lines 311-313).

19. LN 562 NCVs were also detected at contig levels based on the context. Are these contigs used for the binning in method section “Nucleocytoviricota viruses” or all contigs used for binning. Please refine the methodology.

The binning was performed separately using “all assembled scaffolds” (line 704). Bins containing sufficient NCV phylogenetic markers were further processed (see Methods section “Nucleocytoviricota viruses”). We clarified the section the reviewer refers to: “scaffold viral taxonomy was overlaid. In the sections below we detail the target identification of

Crassvirales, Nucleocytoviricota and Preplasmiviricota (PLVs and virophages), while for other viral sequences we used the Cenote-taker2 taxonomy” (lines 622-624).

Minor concerns:

20. LN 71: The 75% novel viral sequences is higher than the ~30% unique populations in the Antarctic as shown in the GOV paper (<https://pubmed.ncbi.nlm.nih.gov/31031001/>). This is expected because the SO samples were scarce. A species rarefaction curve should be included to tell if the viral richness has reached a plateau.

We have done a sample rarefaction curve and a species accumulation curve that shows that we have sufficiently sequenced the samples deeply enough and sampled our study area (we added these as Fig. R5 and R6). More studies targeting the open Southern Ocean, other coastal regions, under ice, deep ocean and other Antarctic seas are needed to truly estimate the viral diversity in this region. We have added the sample rarefaction curve and a species accumulation curve to the Supplementary Fig. 2. We added the following to the Results: “The sample rarefaction and the species accumulation curves show that we have sequenced deeply enough and with enough sample coverage of the site (Supplementary Fig. 2a-b)” (lines 71-73).

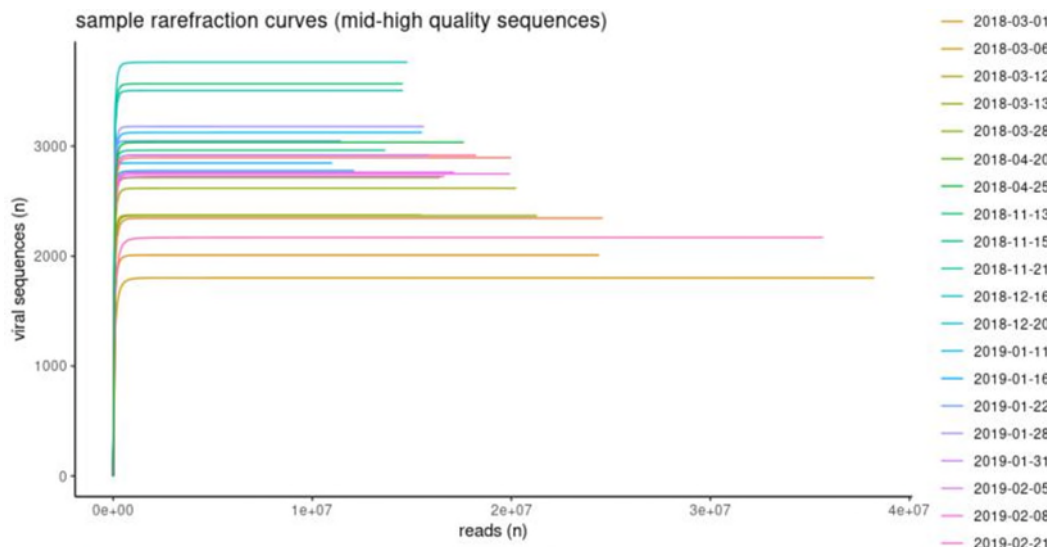


Figure R5 – Sample rarefaction curves for this study.

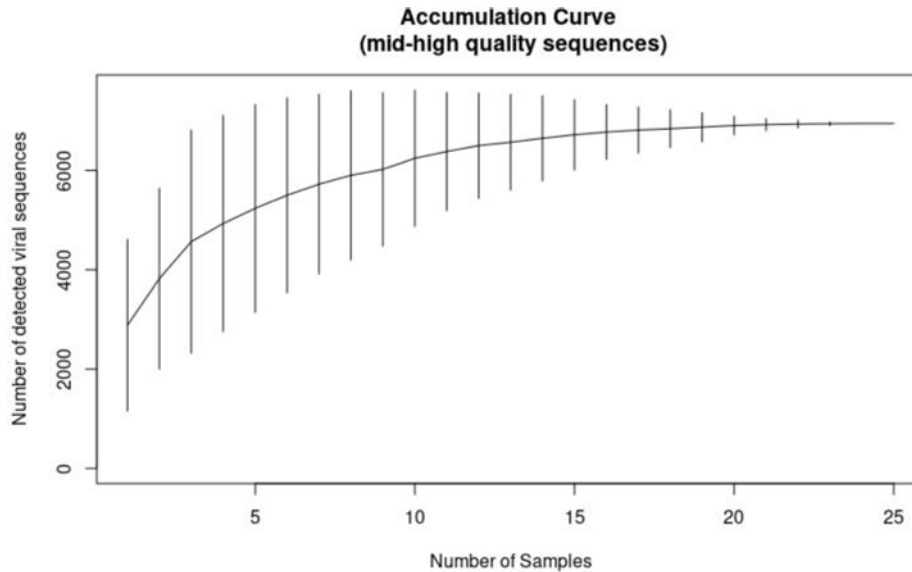


Figure R6 – species accumulation curve for Antarctic viral sequences from this study.

21. Please define early, mid, and late summer with specific months and mark these periods in the plots for clarity.

We have added this to all the main figures.

22. LN 68: Please define the 35%. What about the ratio of reads could be mapped?

We have added this information to the text: "The viral reads represented 59% of the viral fraction metagenomes on average and only 4% of the reads did not map to the co-assembly (Supplementary Fig. 1a)" (lines 70-71).

23. LN 147 Fig.1F is not for archaeal viruses. Should be

Fig1E. *We have made this change.*

0. LN 199 Fig.S8 should be Fig.S7, and the legend of Fig.S7 is missing. *We have made the change (line 220).*

1. It seems the information between Fig.S9 (v-Contact2) and Fig.1 are redundant. Same to Fig.2C and Fig.3C (Viral particle counts); Fig.5C and Fig.1D.

- Fig S9 and Fig.1 show a different set of sequences. Fig S9 is for all scaffolds bigger than 5kb and illustrates the phylum level clustering used to annotate unknown sequences. Fig 1a only has the mid to high quality sequences > 10kb or at least 70% complete according the checkV. (see Methods line 629). Figure legends clarify this as well.

- Fig. 2c and the dots of Fig 3c indeed show viral counts. We added the dotted line as it serves to illustrate the correlation, we think it is still useful for the reader to have it highlighted in Fig. 3c.

- Fig. 5C and 1D show different information. In Fig. 1d we have read abundance at the phylum/class level for NCVs in relation to other Eukaryotic viruses and in Fig 5c we go into family level for NCVs in vertical coverage (we have clarified this in reviewer comment 9).

24. Please use NCVs instead of NCLDVs in Fig.6C.

We have made this change.

25. LN 617 The percentage is not clear. Are multiple copy of TerL, MCP and portal within one contig counted multiple times or once?

Added "no gene duplicates were detected" for clarity.

26. LN 623 It seems that the scaffolds "out of" all ICTV-approved family clades are novel sequences.

Removed "novel" (line 692).

27. It was difficult read the supplementary table 2. Please considering use one spreadsheet to include descriptions of the column headers for all tables.

We have made this change.

28. Please provide the taxonomic lineage information for each accession in supplementary table 3.

We have added the Phylum level taxonomy for each accession (now Supplementary Table 5).

29. The supplementary table 4 seem to be wrong. There are only 17 reference genomes in it and they don't cover the full diversity of NCVs.

We have made a new more complete Supplementary Table (now Supplementary Table 2) that covers all used references and additional information on these study's sequences.

0. I will appreciate authors provide metadata for samples.

The sample metadata is now also provided in Supplementary Table 4.

References:

[1] Moniruzzaman, Mohammad, et al. "Widespread endogenization of giant viruses shapes genomes of green algae." *Nature* 588.7836 (2020): 141-145.

[2] Zhao, Hongda, et al. "A 1.5-Mb continuous endogenous viral region in the arbuscular mycorrhizal fungus *Rhizophagus irregularis*." *Virus Evolution* 9.2 (2023): vead064.

[3] Fischer, Matthias G. "The virophage family Lavidaviridae." *Current issues in molecular biology* 40.1 (2021): 1-24.

- [4] Koslová, Anna, et al. "Endogenous virophages are active and mitigate giant virus infection in the marine protist *Cafeteria burkhardae*." *Proceedings of the National Academy of Sciences* 121.11 (2024): e2314606121.
- [5] Aylward, Frank O., et al. "A phylogenomic framework for charting the diversity and evolution of giant viruses." *PLoS Biology* 19.10 (2021): e3001430.
- [6] Gaïa, Morgan, et al. "Mirusviruses link herpesviruses to giant viruses." *Nature* 616.7958 (2023): 783-789.
- [7] Shkorporov Andrey N., et al. "Create one new order (Crassvirales) including four new families, ten new subfamilies, 42 new genera and 73 new species (Caudoviricetes)" ICTV 2021.022B.R.Crassvirales: <https://ictv.global/ictv/proposals/2021.022B.R.Crassvirales.zip>
- [8] Roux, Simon, et al. "Updated virophage taxonomy and distinction from polinton-like viruses." *Biomolecules* 13.2 (2023): 204.
- [9] Bellas, Christopher, et al. "Large-scale invasion of unicellular eukaryotic genomes by integrating DNA viruses." *Proceedings of the National Academy of Sciences* 120.16 (2023): e2300465120.
- [10] Yutin, Natalya, Didier Raoult, and Eugene V. Koonin. "Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies." *Virology journal* 10 (2013): 1-15.
- [11] Roux, Simon, et al. "Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics." *Nature communications* 8.1 (2017): 858.
- [12] Zolfo, Moreno, et al. "Detecting contamination in viromes using ViromeQC." *Nature biotechnology* 37.12 (2019): 1408-1412.

Dear Reviewers,

We sincerely appreciate the constructive and positive comments you provided during both review rounds. These have substantially enhanced the analysis and overall quality of the manuscript.

To tackle the concerns raised by Reviewer 3, we have thoroughly addressed the issue of contamination by including a collection of eukaryotic and other non-Nucleocytoviricota genomes in our analysis. Additionally, we have included the recently described Mriyaviricetes class, which turned out to be the class of many of the bins previously classified as Momono-like viruses.

A detailed response point-by-point follows below. We are confident that these revisions have significantly strengthened the robustness of our analysis and the reliability of our results. We look forward to your response.

Thank you once again for your time and consideration.

Yours sincerely,

Goncalo Piedade and Prof. Dr. Corina Brussaard (on behalf of the coauthors)

Reviewer #3 (Remarks to the Author):

I'd like to thank authors for the effort in addressing my comments. Most of the responses are adequate and directly address my concerns. The authors have made changes where feasible and provided clear explanations. The improvements, such as the partial Mantel test and adding support values, make the results easier to interpret. The adjustments to the manuscript have made the data analysis more convincing. Overall, the manuscript is much improved. Good luck with the data mining on crassphages in the large database.

I still have few comments on the NCV and contamination.

1. Concerning comment #3 by reviewer #1, about using additional tools to define the viral genomes and gaps: I noticed that the authors used ViralRecall to decontaminate the cellular sequences. However, the ViralRecall scores, which have the function to define the viral regions by considering the penalty from cellular signals, were not provided in the main text or Table S2. Including the ViralRecall scores could help define the NCV regions and thus address the concerns.

We have added “removing those with score < 0” in line 660 of the revised manuscript.

2. Related to the comment above and my original comment #6: "The high detection ratio of Mamonoviridae with NCVs is abnormal...". I selected the RNAPL (GVOGm0023) sequence of 2-643266.cc.b31 (the longest bin) and did a quick blastP analysis. The top hits in the NCBI RefSeq and NR databases are all from cellular organisms, with identity ranges from 37%-50% (RefSeq) and 49%-59% (NR). When I restricted the search to the Nucleocytoviricota (taxid:2732007) in the NR database, it yielded hits with identity ranges from 26%-37%. Although this may be due to incorrect labels in the NCBI databases, and the blastP identity doesn't necessarily represent evolutionary relatedness, I am unable to judge whether those mamonovirus markers are viral, or they are originally cellular genes. We know that viral RNAPs resemble eukaryotic RNAPs (I and II) [DOI: 10.1073/pnas.1912006116]. Therefore, an HMM-based detection may yield false positives on some eukaryotic marker genes, leading to false positives in giant virus bin detection, especially given that authors used two marker genes as the cut-off for this clade.

We thank the Reviewer for reframing this point and putting us on the right track. We have run separate phylogenetic analysis per marker gene including a comprehensive set of Nucleocytoviricota (NCVs), Mirusviricota, Herpesvirales, Eukaryotic and Archaeal genomes [1], including the recently described NCV class Mriyaviricetes [2]. This has allowed us to identify 6 bins previously classified as “Momono-like viruses” that were mostly composed of Eukaryotic-like RNAPol, DNAPolB and, and DNA topoisomerase II. We have adapted the methods section accordingly:

“Separate phylogenetic trees for the markers were built by combining the bins with a comprehensive set of NCVs, Mirusviricota, Herpesvirales, Eukaryotic and Archaeal genomes collected by Karki et al. 2024 (ref¹⁰⁷), additionally including the recently described NCV class Mriyaviricetes⁴⁷. We further removed scaffolds containing markers falling within the Eukaryotic clades for RNA polymerase, DNA topoisomerase I, and DNA polymerase B. The phylogenetic trees

were built by detecting and aligning the seven genetic markers⁴⁶ using *nclsv_markersearch*. The alignments were trimmed of regions where >20% of the sequences have a gap and with an entropy score below 0.55 using *BMGE* v1.12111. The trimmed alignments were used to build a maximum likelihood tree using *iqtree2*108 the best model finder option (-m MFP) with ultrafast bootstrap of 1,000 replicates." (lines 659-666)

We also decided to add to the analysis references belonging to the recently described NCV candidate class *Mriyaviricetes* [2]. Interestingly, we found that the *VLTF3*, *ATPase* and *MCP* fall within the candidate *Gamadviridae* sequences for the remaining viruses previously classified as "Momonoviridae". Given these viruses are small (35–45 kb) and have been found within genome assemblies [2,3], we decided to remove all scaffolds that did not contain a *Mriyaviricetes* *VLTF3*, *ATPase* and *MCP*. We have adapted the Methods:
"The *Mriyavirus* sequences belonged to large genomic bins which also contained all previously detected Eukaryotic-like RNA polymerase and DNA polymerase B genes. We removed this Eukaryotic host contamination by retaining only the contigs containing the marker genes *VLTF3*, *A32* and *MCP*⁴⁷." (line 672-676)

In light of the new result, we have changed the Results and Discussion to reflect these new finding:

- candidate class *Mriyaviricetes*⁴⁷ (n=9). (line 204)
- The nine *Mriyaviruses* all belong to the candidate family *Gamadviridae* and were found inserted into larger bins recovered from the cellular fraction. Eight of these bins had high Eukaryotic content, representing five unclassified Eukaryotes and four classified as *Phaeocystis antarctica* (Supplementary Table 2). We hypothesize that these viruses were binned together with their host, representing temperate or persistent infection (Supplementary Table 2), consistent with their presence inside the genome assemblies of *Phaeocystis*⁵¹ and other eukaryotes⁴⁷. (lines 222-227)
- Overall, this study provides a comprehensive characterization of the *Bamfordvirae* kingdom diversity at the study site. This characterization expands the known PLV and viroplasm diversity, includes the recovery of novel NCVs together with those belonging to the candidate class *Mriyaviricetes*⁴⁷, and also identifies members of the recently described candidate phylum *Mirusviricota*⁴⁸. (lines 328-332)

3. Some taxa should be in italic: LN244 "*Mamonoviridae*," LN245 "*Mirusviricota*," the legend in Fig 5A, and LN909 "*Nucleocytoviricota*" and "*Preplasmiviricota*."

We have changed accordingly.

4. LN235. Should "late and early season" be changed to "...early summer.."? I see that there are six genomes represent an "early summer only" seasonality pattern, but I don't understand what "late" stands for.

We have changed “late and early season” to “autumn and early summer” (line 198). Have similarly changed in lines 220 and 270.

References:

[1] Karki, S., Barth, Z. K., & Aylward, F. O. N. (2024). *Chimeric Origin of Eukaryotes from Asgard Archaea and Ancestral Giant Viruses*. *bioRxiv*, 2024-04.

[2] Yutin, N., Mutz, P., Krupovic, M., & Koonin, E. V. (2024). *Mriyaviruses: small relatives of giant viruses*. *Mbio*, e01035-24.

[3] Roitman, S., Rozenberg, A., Lavy, T., Brussaard, C. P., Kleifeld, O., & Béjà, O. (2023). *Isolation and infection cycle of a polinton-like virus virophage in an abundant marine alga*. *Nature Microbiology*, 8(2), 332-346.

Dear Reviewer,

We sincerely appreciate the constructive and positive comments you provided during the review process. We are glad to hear the changes address all your concerns, we agree these have improved the robustness of analysis and the manuscript. We have also removed the duplicated reference.

Thank you once again for your time and consideration.

Yours sincerely,

Goncalo Piedade and Prof. Dr. Corina Brussaard (on behalf of the coauthors)

REVIEWERS' COMMENTS

Reviewer #3 (Remarks to the Author):

I would like to thank the authors for addressing my comments. I'm pleased that my suggestions regarding the identification of novel virus-like sequences (Mriyaviricetes) were helpful. The manuscript has been much improved, and I have no further concerns.

One small detail: it seems that a reference may be duplicated (#101 and #108).