

Biogeography and host interactions of CPR and DPANN viruses in acid mine drainage sediments

Corresponding Author: Professor Linan Huang

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)

In the manuscript entitled: Biogeography and host interactions of CPR and DPANN viruses in acid mine drainage sediments across southeast China, Lin et al conducted a study based on metagenomics data that were acquired using 90 samples from acid mine drainage, specifically focusing on CPR and DPANN viruses, their hosts and the symbionts' hosts. Regarding viruses, the dataset consisted of 241 CPR-associated vOTUs and 97 DPANN-associated vOTUs, out of 5,678 vOTUs. In a previous publication (Gao et al, 2022, Nature communications), the authors used the same dataset, providing a global analysis not specifically focused on DPANN or CPR, therefore the new study represents a different and invaluable analysis. In addition, in the present study, the authors newly conducted a metatranscriptomic analysis using more recently-collected samples.

As far as I can judge, the study is well conducted with accurate and thorough bioinformatics and biostatistics analyses. It was however more difficult for me to review the last result section, on "Interactions between viruses and symbiotic hosts of DPANN/CPR", as I am not familiar with the precise statistical tests used in this section. The study brings original and very significant insight into the diversity of viruses infecting DPANN and CPR, which are still poorly known, as well as notable results and hypotheses regarding their ecological interactions and role. In particular, the authors show complementary trends regarding the niches, with DPANN and their viruses being more abundant in the most extreme conditions (low pH, high electrical conductivity), while CPR and their viruses show the opposite pattern. The dominant life style of viruses also differs: the proportion of vOTUs predicted to be temperate was much higher for the CPR viruses than for the DPANN ones. Moreover, the host-virus ratios and lysogeny rates showed a link with major environmental stresses, notably the electrical conductivity, again showing distinct trends for CPR and DPANN. Analysis of the Auxiliary Metabolic Genes suggested that some viruses possibly assist their hosts in resistance to stress and other functions. Metatranscriptomic analysis revealed that several viruses were actively transcribed. Finally, the authors suggested that interactions could exist between viruses and hosts' symbionts, that some hosts' symbionts may protect DPANN/CPR symbionts by suppressing viral replication. I have some comments on specific points and therefore suggest that the manuscript is accepted for publication with minor revisions.

Main comments

1/ Although the hypothesis that viruses assist their host in resistance to stress is invaluable and makes sense, for some of the most abundant AMG (e.g. queE, cysH), I feel that additional hypotheses may also exist, that these genes could play a more general role in the virus life cycle, which may actually explain their abundance. These additional hypotheses should usefully also be mentioned in the main text of the manuscript.

Regarding cysH, it was identified as an abundant AMG in many anoxic environments and Wang et al, in Potential metabolic and genetic interaction among viruses, methanogen and methanotrophic archaea, and their syntrophic partners, ISME Communications, <https://doi.org/10.1038/s43705-022-00135-2>, suggested that cysH could be used by viruses to "synthesize more cysteine, which is the only amino acid that can form disulfide bonds that stabilize viral protein structure", through enhanced assimilatory sulfate reduction pathway.

Regarding queE, Hutinet et al, in 7-Deazaguanine modifications protect phage DNA from host restriction systems. Nat Commun <https://doi.org/10.1038/s41467-019-13384-y>, suggested that FolE, QueE, QueD and QueC, could play a role in DNA modification that protect phage DNA from host restriction systems.

Finally, PhoH has been identified as abundant in other environments, such as the sea (Goldsmith et al, in Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity. Appl Environ Microbiol <https://doi.org/10.1128/AEM.05531-11>) suggesting a role that is not necessarily specific to acid mine drainage, but that is still

worth noting.

2/ For the metatranscriptomics analysis, ribodepletion is not mentioned in the material and methods section. It is indicated that "Contaminated rRNA sequences were identified and eliminated". Could the author confirm that no ribodepletion was performed and if so, state it more explicitly in the text and explain why?

3/ In the supplementary materials, in Supplementary Fig. 6, the authors show the correlation of viral composition (PC1) and environmental factors based on PCoA analysis. In the main text of the manuscript, in Fig 2e and 2f, the authors provide a visualization based on NMDS. Would it be possible to provide, in the supplementary materials, a similar visualization but based on PCoA as well, to have a more complete and consistent view?

4/ From the material and method section, I understand that the viral genomes are not necessarily complete, but that they are longer than 10kb, is this correct? If so, I think that it should be stated more explicitly that the viral genomes are not necessarily complete genomes. Indeed, in such case, the proportion of temperate viruses has high chances to be underestimated, due to missing integrase genes for example, so that the true proportion of virulent viruses in DPANN could be lower than the observed one, although still higher than in CPR viruses.

5/ I feel that some sentences in the abstract could be more nuanced and should be completed with additional information, for example, lines 35-38:

"DPANN and their viruses colonize more extreme conditions through Piggyback-the-host dynamics, while CPR viruses adopt Kill-the-host strategies under acidity and salinity stresses, suggesting a role of viral predation in the niche partitioning between CPR and DPANN along environmental gradients."

Could be modified as:

"DPANN and their viruses colonize more extreme conditions mostly through Piggyback-the-winner-host dynamics, while CPR viruses more frequently adopt Kill-the-host strategies under acidity and salinity stresses, suggesting a role of viral predation in the niche partitioning between CPR and DPANN along environmental gradients."

(The word winner was missing.)

Moreover, this sentence alone could give the impression that lysogeny is more frequent in DPANN viruses than in CPR viruses, whereas the opposite is true on average (Figure 1e): maybe this should also be mentioned in the abstract.

Similarly, on lines (459-461), it is written: "Of note, the detected CPR viruses shift toward a Kill-the-host strategy under intensified environmental stresses, where DPANN and their viruses tend to adopt a relatively mutualistic dynamic." which can be ambiguous again, considering the dominant virulent life style of DPANN viruses observed in this study. This sentence could be revised as; "Of note, the detected CPR viruses shift toward a Kill-the-host strategy under intensified environmental stresses, where DPANN show an opposite trend, with an increase of temperate life style under intensified environmental stresses" (or similar).

6/ For the viral genomes shown in Figure 4b, it could be useful to check that the coverage obtained through read mapping is roughly homogeneous along the genome, to assess whether the viral genome could be chimeric. Have the authors done so?

Minor modifications

1/ Line191: "MAT and stresses". This is the first occurrence of MAT, which should be defined here (mean annual temperature) instead of on line 495.

2/ Lines 300-301 "These enzymes have previously been identified in cellular organisms and giant viruses previously" One occurrence of the word "previously" should be deleted.

3/ Line 333 and followings: "A total of 780 MAGs were identified as the SHs, of which 120 were annotated as sulfate-reducing microorganisms (...). It could be interesting to indicate here the total number of MAGs to get an idea of the proportion of SHs.

4/ Line 553, it is stated: "only the AMGs flanked by two viral genes were retained". However, when looking at Figure 4b, not all the AMGs seem to be flanked by 2 viral genes (some are labeled as "unannotated" or "other". Could the authors please clarify this point.

5/ Figure 1a: it would be useful to indicate in the legend that the region shown is south east China, and to add a km scale

6/ Figure 3: in the y axis legend, vh could be replaced by vhr for more clarity.

7/ Figure 5b, legend: "infected phage" should be replaced by "infecting phage"?

8/ In the supplementary Figure 6b, it would be useful to label the groups (from left to right, as far as I could guess: A, B, C, D, E, F).

9/ In the supplementary Figure 19a, in "a correlation abalysis", replace the b with an n

10/ In the supplementary data, it could be useful to place S11 as the first sheet (S1), to have an overview of the names of the

sites and samples before focusing on the data.

Reviewer #2

(Remarks to the Author)

General Comments

In this manuscript Lin et. al examine the viral communities associated with CPR Bacteria and DPANN Archaea in acid mine drainage systems across a broad geographical range. In general I find the paper is well written and the underlying data is interesting with some curious findings, particularly regarding accessory metabolic genes identified in the viruses of these symbiotic organisms. However, I have several concerns regarding under-described methodology and conclusions that are not fully supported by the data in my opinion. From a methods perspective the description of viral lifestyle annotation and host species identification as well as calculation of lysogeny rate need better description. With regards to conclusions I am unconvinced that viruses are the ones driving niche differentiation and not the inverse. I am also sceptical of the inferences from the regression lines of VHR and lysogeny rates due to generally low R2 values and high data spread. Further, in my view the "putative host" analyses are not convincing enough to justify referring to the species as "putative hosts" since for the majority of CPR and DPANN several hosts across multiple phyla and often domains are predicted and no known CPR or DPANN has such a broad range.

In summary, whilst the quality of the underlying data seems good I believe major revisions are necessary to clarify some of the methodological concerns I have and some sections require rephrasing to better represent what the data shows.

Major Concerns:

Data that should be added

There are a couple of instances in the manuscript where the Supplementary Data file provided does not include underlying data from which analyses were performed. In one example I was able to find these in the Figure production zip but these data should be included in the Supplementary Data file so that readers have access to this easily. This include the metadata file that contains chemical data for the different samples and a table containing abundance estimates for the "putative host" species.

Classification of Viruses as lytic vs lysogenic

The authors use a combination of several different tools as well as manual assessment to assign viruses as either lytic or lysogenic. However, it is not well described how these were combined to make the determination of lifestyle. I note that the methods state "A total of 44 consistent annotation results were obtained along with five ambiguous results, which were manually rectified" but Supplementary Data 1 assigns 230 viruses as lysogenic. Most viruses in this table do not appear to have agreement across all tools used for assessing lysogenic vs lytic based on the data presented in the table, for example BP4_virus117 which seems to be predicted as lysogenic solely from DeePhage? How did the authors combine the tool outputs to assign lifestyle? Were all viruses included in downstream analyses or only those with confident lifestyle assignments? I note that in the files provided for production of figures there are only two factors listed for lifestyle annotations (DeePhage and pfam) and seven examples where these conflict and one was selected over the other with no explanation why.

It should also be made more clear if the Life Strategy column of Supplementary Data 1 corresponds to direct output from DeePhage or a combined assessment based on all tools used, at present I assume it is solely DeePhage since this output is not provided anywhere else. The authors should also edit their figures and tables to use consistent language regarding viral lifestyle. At present the text refers to lytic vs lysogenic whilst several supplementary figures and all supplementary data refers to temperate vs virulent.

Lysogeny Rate

What is the lysogeny rate presented in various figures representing? The calculation of this value is not described anywhere in the methods as far as I can tell and the term is ambiguous since lysogenic viruses have a lytic phase of their lifecycle depending on conditions. I would normally interpret this phrase as representing how many viruses are actively lysogenic vs lytic but I cannot see how the authors would infer this from their dataset so I assume it instead represents the ratio of predicted lysogenic vs lytic viruses, which from the supplied R script appears to be the case. This should be explained in the methods.

Statistical Analyses of VHR and Lysogeny Rate Variations

I am sceptical as to some of the conclusions being drawn from the calculated lysogeny rates, in particular statements that are made regarding differences in virus lifecycle between CPR and DPANN across pH and EC. The authors draw conclusions regarding correlations between EC and VHR but the R2 of the models used for these inferences are low and the dataset being analysed seems to only include eight samples with an EC > 5 and only seven with a pH > 4, though it's difficult to tell since the authors do not include a table detailing the environmental parameters for each sample. This should be added. The fact that the sample density at higher EC and pH values is so low and the spread of data points in the lower range is so wide combined with the low R2 suggests that the model fit is poor. The methods section also does not clarify if appropriate tests were done prior to model fitting, for example whether the residuals are normally distributed. Looking at the provided R script for this analysis I do not see evidence that this was performed before the model fitting was carried out. For the authors to draw conclusions from statistical analysis of their data in this way they need to provide more details regarding how the model fitting was done to justify their inferences regarding correlations predicted from these models.

As a follow up to this the Authors discuss that the lysogeny rates suggest piggyback-the-winner dynamics but also in the abstract that CPR viruses adopt kill-the-host dynamics at low pH. However, as far as I can see there is no clear indication of this in the data. Piggyback-the-winner implies that lysogeny increases as host abundance and growth is high, in contrast at low growth rates and abundances (such as CPR in low pH environments) lytic lifestyles should be preferred. I do not see

how the VHR values presented here imply a switch away from piggyback-the-winner in CPR viruses at low pH, the data presented are consistent with the predictions of the piggyback-the-winner model.

DPANN and CPR Host Prediction Method

The authors rely upon metabolic interaction predictions and abundance correlations to predict a sweeping range of potential DPANN and CPR hosts. This section needs rephrasing since as far as I can see from the data presented in Supplementary Data 7 the authors have no direct evidence of interactions between the predicted partners. Additionally, virtually every symbiont genome has multiple predicted host species with many having >10 including candidates from either Archaea or Bacteria and multiple phyla as well. As far as I'm aware no DPANN or CPR has ever been shown to have a host range anywhere near as broad as this data implies. What is shown in the table is better represented as positively correlated community members, of which some may function directly as hosts but most are likely not hosts of the CPR and DPANN in question. The section should be rephrased to make this clear.

It is also difficult to assess the significance of the correlations reported in this table without the underlying abundance of the putative host species. The authors provide these in the Supplementary Data for the CPR and DPANN but not for any of the other MAGs mentioned here. Abundance estimates for all the MAGs included in this analysis should be provided.

Identification of virus host identity

The authors describe using seven different methods for assigning viral genomes to putative hosts but do not explain how they combined the outputs of these. In Supplementary Data 2 it seems that the authors accept any host-virus pair prediction as long as it is recovered from at least one of the methods used. Is this the case? If so it should be stated more clearly in the methods section.

Inference of a causal link between viral behaviour and niche differentiation

The authors at several points make the argument that differences in their observations of viral lifestyle between CPR and DPANN play a causal role in niche differentiation of their host species. I see no evidence presented in this manuscript that supports a causal relationship in this direction. It could just as easily, and I think more convincingly, be argued that the niche differentiation of the two symbiotic lineages has a causal role in favouring the different patterns of viral lifestyle observed. If the authors wish to argue that viral predation patterns are partially responsible for the different patterns observed for DPANN vs CPR with regards to environmental conditions then they need to provide a much more convincing argument for how this is supported by their data.

Minor Concerns:

Use of Phyre2 vs AlphaFold2

I am curious why the authors opted to use Phyre2 for structure prediction as opposed to the generally more accurate AlphaFold2 option? The results seem robust regardless but I would have expected AlphaFold2 to be a better option for this.

Specific Comments:

Ln 53: "and constitute an essential component of the tree of life"

Not clear what this means.

Ln 57: Several references that could be cited here:

Wurch 2016, Golyshina 2017, Hamm 2019, St John 2019, La Cono 2020, Batinovic 2021, Kato 2022, Xie 2022, Kuroda 2022a, Kuroda 2022b, Yakimov 2022

If the authors wish to only cite acidophile works then as a minimum Golyshina 2017 PMID: 28680072 should be added.

Ln 62: "consequently exert significant impacts on prokaryotic community"

Language expression is incorrect here, they have impacts or they exert pressure, either is fine.

Ln 67: "and CPR lineages17-20"

Also Crits-Cristoph 2016

Ln 119 – 122: Since this study does not include any isolation or cultivation experiments this should be rephrased as "Predicted to adopt lysogenic/lytic lifecycles". As written the sentence implies the lifecycles of the relevant viruses has been directly observed which is not the case.

Ln 138 – 140: As currently written the authors present extreme conditions as "stressful" generally. However, this is entirely dependent on the frame of reference used, extremophiles typically do not exhibit stress markers under the conditions they inhabit (in this case low pH) and thus for the DPANN found in these low pH environments the conditions are likely not "stressful". I would recommend using an alternative to "stress" such as "more extreme environmental conditions" which is agnostic to the stress state of the organism.

Ln 391: Genome streamlining is not unique to CPR and DPANN, it is virtually ubiquitous amongst them but it is definitely not unique to them.

Ln 436 – 437: This sentence needs to be rephrased, at present it implies CPR and DPANN evolutionarily diverged from one another which is not accurate. I assume the authors are referring to the dynamics of the evolutionary trajectory of the two lineages and not actual divergence of the two but this should be rephrased to clarify.

Ln 445 – 446: The rates of evolution of CPR and DPANN are not "attributed" to gene loss and acquisition. Both of these are connected to a host dependent lifestyle which results in relaxation of selective pressure on otherwise core functions allowing

increased rates of both mutation and total gene loss.

Ln 450 – 451: This needs rephrasing, DPANN are not closely related to the Asgard archaea from which Eukaryotes emerged and the authors have not shown a direct evolutionary link between the viruses of DPANN and those of Eukaryotes. This statement should be modified to reflect that the author's data suggest acquisition of these characteristics by viruses may predate the divergence of Eukaryotes from Archaea but that there is not a direct evolutionary link between the viruses presented in this study and eukaryotic viruses with these characteristics.

Ln 457 – 459: I see no evidence in the reference provided that CPR genomes are relatively more streamlined than DPANN genomes nor do I see that the original authors made this claim. Could the authors clarify what they are basing this claim on?

Ln 515-516: This sentence is unclear, are the authors saying that if the viral genome is aligned to the host genome then it is classified as a prophage?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Dear Editor, dear Authors,

I feel that the reviewers' comments have been rigorously and appropriately addressed, and that the authors are now more cautious in their conclusions, avoiding any over-interpretation of the data. I am in favour of accepting the manuscript for publication.

(Remarks on code availability)

I did not run the scripts, but I saw that they are clear and well structured, that the packages used are generally well known in the field of microbial ecology or for R users, and that a readme file explains the purpose of each script, indicating the nature of the analysis and which figure(s) in the article the script corresponds to, if relevant.

Reviewer #2

(Remarks to the Author)

In this revision Lin et. al. have provided a number of additional/alternative analyses to address my comments from the original submission. The authors have replaced the original statistical analyses with more robust options better capable of handling the dataset and retrieved consistent results with greater support. They have clarified all the methodological questions I raised and provided descriptions of these in text in addition to providing all the underlying data I requested. The additional analyses regarding assignment of DPANN and CPR hosts was not requested but has significantly improved this section alongside. The authors have also addressed all phrasing and language issues I raised with the previous version of the manuscript.

I have no further comments and am supportive of publication.

(Remarks on code availability)

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/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In the manuscript entitled: Biogeography and host interactions of CPR and DPANN viruses in acid mine drainage sediments across southeast China, Lin et al conducted a study based on metagenomics data that were acquired using 90 samples from acid mine drainage, specifically focusing on CPR and DPANN viruses, their hosts and the symbionts'hosts. Regarding viruses, the dataset consisted of 241 CPR-associated vOTUs and 97 DPANN-associated vOTUs, out of 5,678 vOTUs. In a previous publication (Gao et al, 2022, Nature communications), the authors used the same dataset, providing a global analysis not specifically focused on DPANN or CPR, therefore the new study represents a different and invaluable analysis. In addition, in the present study, the authors newly conducted a metatranscriptomic analysis using more recently-collected samples.

As far as I can judge, the study is well conducted with accurate and thorough bioinformatics and biostatistics analyses. It was however more difficult for me to review the last result section, on "Interactions between viruses and symbiotic hosts of DPANN/CPR", as I am not familiar with the precise statistical tests used in this section. The study brings original and very significant insight into the diversity of viruses infecting DPANN and CPR, which are still poorly known, as well as notable results and hypotheses regarding their ecological interactions and role. In particular, the authors show complementary trends regarding the niches, with DPANN and their viruses being more abundant in the most extreme conditions (low pH, high electrical conductivity), while CPR and their viruses show the opposite pattern. The dominant life style of viruses also differs: the proportion of vOTUs predicted to be temperate was much higher for the CPR viruses than for the DPANN ones. Moreover, the host-virus ratios and lysogeny rates showed a link with major environmental stresses, notably the electrical

conductivity, again showing distinct trends for CPR and DPANN. Analysis of the Auxiliary Metabolic Genes suggested that some viruses possibly assist their hosts in resistance to stress and other functions. Metatranscriptomic analysis revealed that several viruses were actively transcribed. Finally, the authors suggested that interactions could exist between viruses and hosts'symbionts, that some hosts' symbionts may protect DPANN/CPR symbionts by suppressing viral replication.

I have some comments on specific points and therefore suggest that the manuscript is accepted for publication with minor revisions.

Response: Thank you for your positive feedback on our study. We appreciate the opportunity to clarify the statistical methods used, and we hope the following explanation addresses your concern. The core aim of the section “Interactions between viruses and symbiotic hosts of DPANN/CPR” was to move beyond simple co-occurrence and uncover direct and robust associations between DPANN/CPR viruses and the symbiotic hosts (SHs; defined as members in symbiosis with the virus-targeted DPANN/CPR), while rigorously controlling for potential confounding factors (such as the abundance of the primary DPANN/CPR host itself and other community members). To address this, we firstly used a strengthened criteria (suggested by Reviewer #2) to select SHs for each virus-targeted DPANN/CPR population: (i) any HGT-identified candidates, (ii) the candidate with the highest abundance correlation, and (iii) the candidate with the highest metabolic interaction potential (MIP) value within their respective domains (Archaea for DPANN, Bacteria for CPR). Then, we employed a two-tiered statistical approach:

1. Mantel test: This test assesses whether the overall community structure of one group (e.g., viruses) is significantly correlated with another (e.g., SHs) across all samples. The strong, significant correlation (even after accounting for the DPANN/CPR host community) suggests a significant community-level correlation between the viruses and SHs.
2. Partial correlation network (newly added analysis): As described by Fabbri

et al. (Environ Int. 2023;173:107856.), this method identifies independent associations between one member (e.g., one virus) and another member (e.g., one SH) after statistically controlling for all other variables (including the DPANN/CPR host and other co-occurring SHs and viruses). This allows us to infer direct, potential interactions, revealing the positive and negative linkages discussed.

We have revised the Methods section to include a clearer description of this rationale. Other specific comments listed in the attached file have also been meticulously addressed.

Main comments

1/ Although the hypothesis that viruses assist their host in resistance to stress is invaluable and makes sense, for some of the most abundant AMGs (e.g. queE, cysH), I feel that additional hypotheses may also exist, that these genes could play a more general role in the virus life cycle, which may actually explain their abundance. These additional hypotheses should usefully also be mentioned in the main text of the manuscript.

Regarding cysH, it was identified as an abundant AMG in many anoxic environments and Wang et al, in Potential metabolic and genetic interaction among viruses, methanogen and methanotrophic archaea, and their syntrophic partners, ISME Communications, <https://doi.org/10.1038/s43705-022-00135-2>, suggested that cysH could be used by viruses to “synthesize more cysteine, which is the only amino acid that can form disulfide bonds that stabilize viral protein structure”, through enhanced assimilatory sulfate reduction pathway.

Regarding queE, Hutinet et al, in 7-Deazaguanine modifications protect phage DNA from host restriction systems. Nat Commun <https://doi.org/10.1038/s41467-019-13384-y>, suggested that FolE, QueE, QueD and QueC, could play a role in DNA modification that protect phage DNA from host restriction systems.

Finally, PhoH has been identified as abundant in other environments, such as the sea (Goldsmith et al, in Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity. Appl Environ Microbiol <https://doi.org/10.1128/AEM.05531-11>) suggesting a role that is not necessarily specific to acid mine drainage, but that is still worth noting.

Response: We sincerely thank you for this insightful and informative comment. In the revised Results and Discussion sections regarding highly abundant AMGs (such as *queE*, *cysH*, and *PhoH*), we have explicitly acknowledged and briefly addressed alternative hypotheses about their potential roles in core viral processes—including virion stability and DNA protection or modification—alongside the host assistance perspective (Line 237-238, Line 270-271, and Line 407-409). These discussions were concise due to manuscript length constraints, but supported by appropriate citations to relevant literature.

2/ For the metatranscriptomics analysis, ribodepletion is not mentioned in the material and methods section. It is indicated that “Contaminated rRNA sequences were identified and eliminated”. Could the author confirm that no ribodepletion was performed and if so, state it more explicitly in the text and explain why?

Response: A ribosomal RNA (rRNA) depletion step was conducted prior to library construction. This is now stated in the Methods section (Line 602-604): “The rRNA transcripts were removed using Ribo-off rRNA Depletion Kit V2 (Bacteria) (Nanjing Vazyme Biotech Co.,Ltd.) in strict accordance with the provided instructions.” Despite the fact that the experiment eliminated the majority of rRNA, it remains essential to further remove rRNA through bioinformatics approaches. Approximately 30% of the metatranscriptomic sequences were recognized as rRNA sequences and subsequently removed using SortMeRNA software.

3/ In the supplementary materials, in Supplementary Fig. 6, the authors show

the correlation of viral composition (PC1) and environmental factors based on PCoA analysis. In the main text of the manuscript, in Fig 2e and 2f, the authors provide a visualization based on NMDS. Would it be possible to provide, in the supplementary materials, a similar visualization but based on PCoA as well, to have a more complete and consistent view?

Response: Thank you for the suggestion. We have included a new supplementary figure (Supplementary Fig. 6d) showing the PCoA-based distribution of the total viral community composition. This PCoA result is consistent with the patterns observed in Fig. 2e and 2f.

4/ From the material and method section, I understand that the viral genomes are not necessarily complete, but that they are longer than 10kb, is this correct? If so, I think that it should be stated more explicitly that the viral genomes are not necessarily complete genomes. Indeed, in such case, the proportion of temperate viruses has high chances to be underestimated, due to missing integrase genes for example, so that the true proportion of virulent viruses in DPANN could be lower than the observed one, although still higher than in CPR viruses.

Response: According to MIUViG standards (Roux et al., Nat Biotechnol. 2019;37(1):29-37.), viral genome fragments longer than 10kb are acceptable for ecological analysis (e.g., taxonomic classification, diversity estimation). This threshold is also widely adopted in recent environmental omic research (e.g., Gao et al., Nat Commun. 2022;13(1):2389.; Luo et al., Microbiome. 2022;10(1):190.). We have now explicitly stated that contigs >10 kb not necessarily represent complete genomes (Line 530-532). While this may lead to bias in the detection of temperate virus, our prediction of viral lifestyle relied primarily on the deep learning-based tool Deephage, which does not depend solely on marker genes such as integrase, thereby reducing the potential impact of genome fragmentation on our conclusions.

5/ I feel that some sentences in the abstract could be more nuanced and should be completed with additional information, for example, lines 35-38:

"DPANN and their viruses colonize more extreme conditions through Piggyback-the-host dynamics, while CPR viruses adopt Kill-the-host strategies under acidity and salinity stresses, suggesting a role of viral predation in the niche partitioning between CPR and DPANN along environmental gradients."

Could be modified as:

"DPANN and their viruses colonize more extreme conditions mostly through Piggyback-the-winner-host dynamics, while CPR viruses more frequently adopt Kill-the-host strategies under acidity and salinity stresses, suggesting a role of viral predation in the niche partitioning between CPR and DPANN along environmental gradients."

(The word winner was missing.)

Moreover, this sentence alone could give the impression that lysogeny is more frequent in DPANN viruses than in CPR viruses, whereas the opposite is true on average (Figure 1e): maybe this should also be mentioned in the abstract.

Response: The sentence has been revised as: "Our data reveal deterministic processes as the primary driver of virome assembly shaping the distinct distribution patterns of CPR and DPANN viruses. While lifestyle prediction shows higher lytic virus diversity associated with DPANN, both CPR/DPANN viruses likely use the Piggyback-the-winner (PtW) strategy to co-exist with hosts in AMD sediments, with CPR viromes exhibiting increased lysis in low host-density regimes under intensive acidity/salinity conditions."

Similarly, on lines (459-461), it is written: "Of note, the detected CPR viruses shift toward a Kill-the-host strategy under intensified environmental stresses, where DPANN and their viruses tend to adopt a relatively mutualistic dynamic." which can be ambiguous again, considering the dominant virulent life style of

DPANN viruses observed in this study. This sentence could be revised as; "Of note, the detected CPR viruses shift toward a Kill-the-host strategy under intensified environmental stresses, where DPANN show an opposite trend, with an increase of temperate life style under intensified environmental stresses" (or similar).

Response: The sentence has been revised as (Line 480-486): "Of note, both CPR and DPANN viruses tend to adopt PtW strategy in AMD sediments. However, their lifestyle covaried with distinct host distributions under environmental gradients: the detected CPR viruses are predicted to exhibit increased lysis in low host-density regimes under intensive acidity/salinity conditions, while DPANN viruses with higher lytic diversity tend to adopt a relatively mutualistic dynamic."

6/ For the viral genomes shown in Figure 4b, it could be useful to check that the coverage obtained through read mapping is roughly homogeneous along the genome, to assess whether the viral genome could be chimeric. Have the authors done so?

Response: We performed a read mapping and coverage analysis as follows:

1. High-quality sequencing reads were mapped back to the assembled viral genome using Bowtie2.
2. The resulting SAM file was converted to a sorted BAM file.
3. The sorted BAM file was indexed using samtools index.
4. Genome-wide coverage depth was calculated using samtools depth -a.

The results of this analysis showed that the read coverage was uniform across the entire length of the viral genome (Supplementary Fig. 12). We observed no significant, abrupt drops in coverage to zero (which would indicate a large assembly gap) and no sharp, localized peaks or coverage breakpoints.

Minor modifications

1/ Line191: "MAT and stresses". This is the first occurrence of MAT, which should be defined here (mean annual temperature) instead of on line 495.

Response: MAT is now defined on line 192.

2/ Lines 300-301 "These enzymes have previously been identified in cellular organisms and giant viruses previously" One occurrence of the word "previously" should be deleted.

Response: One occurrence of the word "previously" has been deleted (Line 307).

3/ Line 333 and followings: "A total of 780 MAGs were identified as the SHs, of which 120 were annotated as sulfate-reducing microorganisms (...). It could be interesting to indicate here the total number of MAGs to get an idea of the proportion of SHs.

Response: The sentence has been revised to show the total number of MAGs (Line 346-347): "With these stringent criteria, a total of 152 MAGs were identified as SHs in our 2625-MAG dataset."

4/ Line 553, it is stated: "only the AMGs flanked by two viral genes were retained". However, when looking at Figure 4b, not all the AMGs seem to flanked by 2 viral genes (some are labeled as "unannotated" or "other". Could the authors please clarify this point.

Response: As noted, AMG candidates were initially identified using DRAM-v, which assigns an auxiliary score from 1 to 3 to genes that are flanked on both sides by viral genes. According to the DRAM-v documentation, this does not require the AMG to be directly adjacent to viral genes; intervening genes (including those labeled as "unannotated" or "other") are permitted, as long as there is at least one viral gene on each side. To provide full transparency and facilitate review, we have now included a new supplementary table (Supplementary Data 6) that lists the AMGs at Figure 4b, along with their

genomic context and validation details.

5/ Figure 1a: it would be useful to indicate in the legend that the region shown is south east China, and to add a km scale

Response: Thank you for the suggestion. We have added a small inset map in the top-left corner showing the full extent of China. The specific provinces where sampling occurred are highlighted in grey to clearly delineate the broader national context and emphasize that the main map focuses on the southeast China region. A scale bar in kilometers also has been added to the main map to provide a clear reference for distance.

6/ Figure 3: in the y axis legend, vh could be replaced by vhr for more clarity.

Response: The legend is corrected as suggested.

7/ Figure 5b, legend: "infected phage" should be replaced by "infecting phage"?

Response: This is corrected.

8/ In the supplementary Figure 6b, it would be useful to label the groups (from left to right, as far as I could guess: A, B, C, D, E, F).

Response: As correctly guessed, the groups from left to right are A, B, C, D, E, and F. We have now explicitly added these labels to Supplementary Figure 6b to avoid any ambiguity, ensuring consistency with the main text.

9/ In the supplementary Figure 19a, in "a correlation abalysis", replace the b with an n

Response: This is corrected.

10/ In the supplementary data, it could be useful to place S11 as the first sheet (S1), to have an overview of the names of the sites and samples before focusing

on the data.

Response: We have reordered the supplementary tables so that the overview of sites and samples (formerly Supplementary Data 11) is now presented as Supplementary Data 1. All subsequent tables and in-text citations have been updated accordingly to reflect this new order.

Reviewer #2 (Remarks to the Author):

General Comments

In this manuscript Lin et. al examine the viral communities associated with CPR Bacteria and DPANN Archaea in acid mine drainage systems across a broad geographical range. In general I find the paper is well written and the underlying data is interesting with some curious findings, particularly regarding accessory metabolic genes identified in the viruses of these symbiotic organisms. However, I have several concerns regarding under-described methodology and conclusions that are not fully supported by the data in my opinion. From a methods perspective the description of viral lifestyle annotation and host species identification as well as calculation of lysogeny rate need better description. With regards to conclusions I am unconvinced that viruses are the ones driving niche differentiation and not the inverse. I am also sceptical of the inferences from the regression lines of VHR and lysogeny rates due to generally low R^2 values and high data spread. Further, in my view the "putative host" analyses are not convincing enough to justify referring to the species as "putative hosts" since for the majority of CPR and DPANN several hosts across multiple phyla and often domains are predicted and no known CPR or DPANN has such a broad range.

In summary, whilst the quality of the underlying data seems good I believe major revisions are necessary to clarify some of the methodological concerns I have and some sections require rephrasing to better represent what the data shows.

Response: We sincerely appreciate your thoughtful evaluation of our manuscript and the concerns regarding methodological clarity and interpretation robustness. We have implemented comprehensive revisions to address each point as detailed below. Firstly, in the Methods, Supplementary tables and Scripts sections, key information and raw data have been supplemented. Specifically, we have provided environmental parameters (Supplementary Data 1) along with the abundance of putative symbiotic hosts (Supplementary Data 15), facilitating a more straightforward analysis of lifestyles and symbiotic hosts. We have also made it clear that the results are obtained based on bioinformatics methods to avoid over-interpretation. Secondly, we supplemented the hypothesis testing of multiple linear regression and found that the residuals did not conform to a normal distribution. Given that the objective of this section is to investigate the environmental factors associated with the VHR and lysogenic proportion, Robust Linear Regression (RLM) and Aggregated Boosted Tree (ABT) analysis were utilized to re-analyze the impacts of various environmental factors. The results obtained were comparable to those of our previous analyses, yet exhibited enhanced explanatory power (R^2). Further, for the prediction of putative symbiotic hosts (SHs) of DPANN/CPR, we supplemented the analysis of horizontal gene transfer events as additional interaction evidence, and found a strong domain-level specificity (Archaea for DPANN, Bacteria for CPR). Therefore, we used a strengthened criteria to select SHs for each virus-targeted DPANN/CPR population: (i) any HGT-identified candidates, (ii) the candidate with the highest abundance correlation, and (iii) the candidate with the highest metabolic interaction potential (MIP) value within their respective domains. Specifically, the average number of the host phyla associated with each CPR/DPANN genome is less than 2 (Supplementary Fig 23). These alterations are achieved through, yet not confined to, the detailed point-by-point responses to the specific comments presented below.

Major Concerns:

Data that should be added

There are a couple of instances in the manuscript where the Supplementary Data file provided does not include underlying data from which analyses were performed. In one example I was able to find these in the Figure production zip but these data should be included in the Supplementary Data file so that readers have access to this easily. This includes the metadata file that contains chemical data for the different samples and a table containing abundance estimates for the “putative host” species.

Response: Thank you for identifying this critical data accessibility issue. The following key datasets are now directly available in the Supplementary Data section. Specifically, environmental parameters for all samples have been incorporated into Supplementary Data 1. Additionally, the abundance of putative symbiotic hosts has been included as Supplementary Data 15.

Classification of Viruses as lytic vs lysogenic

The authors use a combination of several different tools as well as manual assessment to assign viruses as either lytic or lysogenic. However, it is not well described how these were combined to make the determination of lifestyle. I note that the methods state “A total of 44 consistent annotation results were obtained along with five ambiguous results, which were manually rectified” but Supplementary Data 1 assigns 230 viruses as lysogenic. Most viruses in this table do not appear to have agreement across all tools used for assessing lysogenic vs lytic based on the data presented in the table, for example BP4_virus117 which seems to be predicted as lysogenic solely from DeePhage? How did the authors combine the tool outputs to assign lifestyle? Were all viruses included in downstream analyses or only those with confident lifestyle assignments? I note that in the files provided for production of figures there are only two factors listed for lifestyle annotations (DeePhage and pfam) and seven examples where these conflict and one was selected over the other with

no explanation why.

It should also be made more clear if the Life Strategy column of Supplementary Data 1 corresponds to direct output from DeePhage or a combined assessment based on all tools used, at present I assume it is solely DeePhage since this output is not provided anywhere else. The authors should also edit their figures and tables to use consistent language regarding viral lifestyle. At present the text refers to lytic vs lysogenic whilst several supplementary figures and all supplementary data refers to temperate vs virulent.

Response: The initial viral lifestyle predictions were generated using DeePhage software. Subsequently, we refined these predictions by integrating supplementary lysogeny detection methods, including identification of lysogeny-specific genes through Pfam database analysis and lysogeny signal annotation via CheckV. In cases where discrepancies emerged between the predictions of DeePhage and the results of these lysogeny signal detections, the conflicting classifications were annotated in accordance with the signal detection methods. Of the 65 viruses identified as lysogenic through Pfam/CheckV evidence, 54 were consistent with DeePhage annotations. The remaining 11 discordant cases, in which DeePhage identified the viruses as lytic while Pfam/CheckV provided evidence supporting their lysogenic nature, were conservatively reclassified as lysogenic viruses. This has been explicitly stated in the revised Methods section (Line 564-569). To ensure full transparency and reproducibility, all primary software outputs have been archived in the Fig3.zip data folder on https://github.com/linzhl1/ultrasmall_phage_in_AMD. The detailed methodological implementation is now meticulously documented in the accompanying R script (Fig3-abc.dynamic.R). In the revised manuscript, all lifestyle-dependent analyses (e.g., Fig. 1e, host-virus interaction assessments across environmental gradients) have been re-executed using these refined predictions, with no substantive impact on our original conclusions. To standardize the terminology related to viral lifestyle, the terms 'temperate' and 'virulent' have been consistently replaced by 'lysogenic' and 'lytic', respectively,

in the supplementary figures and all supplementary data.

Lysogeny Rate

What is the lysogeny rate presented in various figures representing? The calculation of this value is not described anywhere in the methods as far as I can tell and the term is ambiguous since lysogenic viruses have a lytic phase of their lifecycle depending on conditions. I would normally interpret this phrase as representing how many viruses are actively lysogenic vs lytic but I cannot see how the authors would infer this from their dataset so I assume it instead represents the ratio of predicted lysogenic vs lytic viruses, which from the supplied R script appears to be the case. This should be explained in the methods.

Response: The “lysogeny rate” has been replaced with “lysogenic proportion,” which represents the relative abundance of lysogenic viruses (Xia et al., ISME J. 2023;17(7):1004-1014.). Corresponding explanations have also been added to the Methods section (Line 571-573).

Statistical Analyses of VHR and Lysogeny Rate Variations

I am sceptical as to some of the conclusions being drawn from the calculated lysogeny rates, in particular statements that are made regarding differences in virus lifecycle between CPR and DPANN across pH and EC. The authors draw conclusions regarding correlations between EC and VHR but the R^2 of the models used for these inferences are low and the dataset being analysed seems to only include eight samples with an EC > 5 and only seven with a pH > 4, though it's difficult to tell since the authors do not include a table detailing the environmental parameters for each sample. This should be added. The fact that the sample density at higher EC and pH values is so low and the spread of data points in the lower range is so wide combined with the low R^2 suggests that the model fit is poor. The methods section also does not clarify if appropriate tests were done prior to model fitting, for example whether the residuals are

normally distributed. Looking at the provided R script for this analysis I do not see evidence that this was performed before the model fitting was carried out. For the authors to draw conclusions from statistical analysis of their data in this way they need to provide more details regarding how the model fitting was done to justify their inferences regarding correlations predicted from these models.

Response: We agree understand the concerns that low sample density at high EC (>5, n=8) and pH (>4, n=7), combined with low R^2 values (0.17) in the original model, could affect the reliability of lifestyle inferences. To address these issues, we first enhanced data transparency by providing the full set of environmental parameters for all samples in Supplementary Data 1. Furthermore, hypothesis tests were done prior to model fitting (using `gvlma` function in the `gvlma` R package), which revealed that the original model residuals do not meet the normal distribution. We then re-analyzed the data using Robust Linear Regression (RLM) (replacing Ordinary Least Square regression). This method employs M-estimators to reduce the influence of outliers and heteroscedasticity, making it particularly suitable for small, non-normally distributed datasets. This approach yielded directionally consistent results with improved explanatory power (Supplementary Fig. 7). To cross-validate findings, we implemented Aggregated Boosted Tree (ABT) analysis (following Liang et al. Nat Commun. 2024;15(1):2827.), which also identified EC as a top predictor of VHR/lysogeny variations (Supplementary Fig. 8). Detailed method description has been revised (Line 686-700) and the R script used for the analysis has also been updated in the Fig3.zip data folder on https://github.com/linzhil1/ultrasmall_phage_in_AMD. Given the limitations of data distribution and the moderate values of R^2 and relative influence, we have rephrased the corresponding results and discussions to interpret EC/pH relationships as covarying trends requiring verification, thereby removing the original overinterpreted claims, and acknowledged the risk of overestimation in the high pH/EC region (Line 168-171, Line 182-183; Line 480-493).

As a follow up to this the Authors discuss that the lysogeny rates suggest piggyback-the-winner dynamics but also in the abstract that CPR viruses adopt kill-the-host dynamics at low pH. However, as far as I can see there is no clear indication of this in the data. Piggyback-the-winner implies that lysogeny increases as host abundance and growth is high, in contrast at low growth rates and abundances (such as CPR in low pH environments) lytic lifestyles should be preferred. I do not see how the VHR values presented here imply a switch away from piggyback-the-winner in CPR viruses at low pH, the data presented are consistent with the predictions of the piggyback-the-winner model.

Response: We agree that our original phrasing inaccurately implied CPR viruses abandon Piggyback-the-winner (PtW) strategy at low pH, when in fact the observed lytic predominance under these conditions directly aligns with PtW predictions. We have now revised the abstract and discussion to explicitly state that the lifestyle differences of CPR and DPANN viruses under low-pH/high-EC conditions do not contradict the PtW dynamics, but rather are distinct cases (Line 34-39, Line 480-486).

DPANN and CPR Host Prediction Method

The authors rely upon metabolic interaction predictions and abundance correlations to predict a sweeping range of potential DPANN and CPR hosts. This section needs rephrasing since as far as I can see from the data presented in Supplementary Data 7 the authors have no direct evidence of interactions between the predicted partners. Additionally, virtually every symbiont genome has multiple predicted host species with many having >10 including candidates from either Archaea or Bacteria and multiple phyla as well. As far as I'm aware no DPANN or CPR has ever been shown to have a host range anywhere near as broad as this data implies. What is shown in the table is better represented as positively correlated community members, of which some may function directly as hosts but most are likely not hosts of the CPR and DPANN in question. The

section should be rephrased to make this clear.

Response: We acknowledge limitations of our previous prediction of potential DPANN and CPR hosts through metabolic complementation and abundance correlations could be misinterpreted as implying direct, confirmed symbiotic relationships over an unrealistically broad range. To provide more direct evidence supporting reliable host relationships, we performed a new horizontal gene transfer (HGT) analysis (method details in Line 644-652). The yielded HGT events identified 12 reliable SHs and demonstrated a strong domain-level specificity of the SHs associated with either DPANN or CPR (Supplementary Data 9). Thus, for each virus-targeted DPANN/CPR, candidates within their respective domains (Archaea for DPANN, Bacteria for CPR) were selected as SHs based on any of the following criteria: HGT evidence, highest abundance correlation, or highest MIP value. Accordingly, the Results section were rephrased in the manuscript (Line 338-350). In the Discussion section, we further reminded the limitation of these computational predictions, emphasizing that further experiments integrating cultivation, high-resolution microscopy, and multi-omic approaches are needed to validate our detected symbiotic interactions (Line 447-453).

It is also difficult to assess the significance of the correlations reported in this table without the underlying abundance of the putative host species. The authors provide these in the Supplementary Data for the CPR and DPANN but not for any of the other MAGs mentioned here. Abundance estimates for all the MAGs included in this analysis should be provided.

Response: We have now added the relative abundance estimates for every MAG included in the co-abundance correlation analysis in the supplementary data (Supplementary Data 14, 15).

Identification of virus host identity

The authors describe using seven different methods for assigning viral genomes to putative hosts but do not explain how they combined the outputs of these. In Supplementary Data 2 it seems that the authors accept any host-virus pair prediction as long as it is recovered from at least one of the methods used. Is this the case? If so it should be stated more clearly in the methods section.

Response: The initial identification of potential host-virus pairs involved accepting a prediction if it was recovered by at least one of the following four primary methods: CRISPR spacer matching, prophage detection within host genomes, genome homology matching, and the integrated machine learning approach. Based on these primary methods that are reliable for genus-level host prediction (Fig 1 in Edwards et al. *FEMS Microbiol Rev.* vol. 40,2 (2016): 258-72. and Fig 3 in Roux et al. *PLoS Biol.* 2023 Apr 21;21(4):e3002083.), viruses associated with CPR and DPANN hosts were identified, and we analyzed their broad community-level ecology. Accordingly, the host-virus pairs listed in Supplementary Data 3 reflect associations recovered by one or more of these methods. To enhance the confidence and accuracy of our predictions for subsequent detailed biological interpretation and analysis, we applied a more stringent criterion. Specifically, for the in-depth analyses presented in the main text (such as the examples in the "Interactions between viruses and symbiotic hosts of DPANN/CPR" section, e.g., MAS4_virus85-YF1.bin123 and YP4_provirus51-YP2.bin85), we required supporting evidence from multiple, distinct prediction methods (Supplementary Data 3, which could include tRNA matching and k-mer similarity analysis in addition to the primary methods). These methodological details have been explicitly stated in the revised Methods section (Line 541-557).

Inference of a causal link between viral behaviour and niche differentiation

The authors at several points make the argument that differences in their

observations of viral lifestyle between CPR and DPANN play a causal role in niche differentiation of their host species. I see no evidence presented in this manuscript that supports a causal relationship in this direction. It could just as easily, and I think more convincingly, be argued that the niche differentiation of the two symbiotic lineages has a causal role in favouring the different patterns of viral lifestyle observed. If the authors wish to argue that viral predation patterns are partially responsible for the different patterns observed for DPANN vs CPR with regards to environmental conditions then they need to provide a much more convincing argument for how this is supported by their data.

Response: We concur that our original argument overattributed causality of viral strategies and have now rephrased the results and discussions that the viral lifestyle was covarying with host abundance under environmental gradients, which may be attributed to the pre-existing host adaptations to niches (Line 182-183, Line 480-486). Considering the regulatory impact of viral predation on the size of the host population, we maintained in our discussion the possibility that viral predation may amplify distribution differences by suppressing CPR host populations in more extreme conditions.

Minor Concerns:

Use of Phyre2 vs AlphaFold2

I am curious why the authors opted to use Phyre2 for structure prediction as opposed to the generally more accurate AlphaFold2 option? The results seem robust regardless but I would have expected AlphaFold2 to be a better option for this.

Response: To leverage the latest advances in protein prediction as suggested, we supplementally generated *ab initio* models using AlphaFold2. The resulting models exhibited high predicted accuracy, with pLDDT scores consistently above 70, a threshold indicative of confident backbone placement. Moreover,

subsequent structural similarity searches (via HHsearch) against the PDB using the AlphaFold2 models yielded functional annotations that were largely consistent with our initial analysis. These results have been integrated into Supplementary Data 5 and the Results section (Line 247-248).

Specific Comments:

Ln 53: "and constitute an essential component of the tree of life"

Not clear what this means.

Response: The original phrasing aimed to convey the extraordinary diversity of CPR bacteria and DPANN archaea within the tree of life. This has been rephased as: "and constitute a remarkable position of the tree of life".

Ln 57: Several references that could be cited here:

Wurch 2016, Golyshina 2017, Hamm 2019, St John 2019, La Cono 2020, Batinovic 2021, Kato 2022, Xie 2022, Kuroda 2022a, Kuroda 2022b, Yakimov 2022

If the authors wish to only cite acidophile works then as a minimum Golyshina 2017 PMID: 28680072 should be added.

Response: Thank you for pointing this out. As suggested, we have now cited the key review by Golyshina 2017 NC, Hamm 2019 PNAS and Batinovic 2021 NM in Line 58 to better acknowledge prior work in this area (References 15-17).

Ln 62: "consequently exert significant impacts on prokaryotic community"

Language expression is incorrect here, they have impacts or they exert pressure, either is fine.

Response: The sentence has been rephased as: "significantly impact the prokaryotic community composition, functions and dynamics".

Ln 67: "and CPR lineages17-20"

Also Crits-Cristoph 2016

Response: The reference now has been cited in Line 68 (References 24).

Ln 119 – 122: Since this study does not include any isolation or cultivation experiments this should be rephrased as "Predicted to adopt lysogenic/lytic lifecycles". As written the sentence implies the lifecycles of the relevant viruses has been directly observed which is not the case.

Response: The sentence has been revised to: "Notably, the viruses infecting the two lineages were predicted to adopt contrasting lifestyles: CPR viruses were primarily lysogenic, while DPANN viruses were largely lytic."

Ln 138 – 140: As currently written the authors present extreme conditions as "stressful" generally. However, this is entirely dependent on the frame of reference used, extremophiles typically do not exhibit stress markers under the conditions they inhabit (in this case low pH) and thus for the DPANN found in these low pH environments the conditions are likely not "stressful". I would recommend using an alternative to "stress" such as "more extreme environmental conditions" which is agnostic to the stress state of the organism.

Response: We agree that describing low pH or high EC as universally "stressful" is inappropriate to the extremophiles since they are evolutionarily adapted to such conditions and may not exhibit physiological stress in their native habitats. The text in the manuscript has been revised as follows: "environmental stress" is modified to "environmental conditions became more extreme", and "stressful conditions" is modified to "more extreme conditions". The other relevant contents in the manuscript have also been modified accordingly (e.g., Line 39, line 404).

Ln 391: Genome streamlining is not unique to CPR and DPANN, it is virtually ubiquitous amongst them but it is definitely not unique to them.

Response: The phrasing has been revised from "unique feature" to "defining

feature" to accurately reflect that extreme genome reduction is a hallmark trait of these ultrasmall prokaryotes while acknowledging its occurrence elsewhere (Line 402).

Ln 436 – 437: This sentence needs to be rephrased, at present it implies CPR and DPANN evolutionarily diverged from one another which is not accurate. I assume the authors are referring to the dynamics of the evolutionary trajectory of the two lineages and not actual divergence of the two but this should be rephrased to clarify.

Response: We acknowledge that the original phrasing could be misinterpreted as suggesting direct evolutionary divergence between DPANN (archaea) and CPR (bacteria), which is phylogenetically inaccurate as they represent independent deep-branching lineages. The new wording states: "The nature of the underlying processes governing the distinct ecological trajectories, as well as the evolutionary dynamics, of DPANN and CPR remains elusive."

Ln 445 – 446: The rates of evolution of CPR and DPANN are not "attributed" to gene loss and acquisition. Both of these are connected to a host dependent lifestyle which results in relaxation of selective pressure on otherwise core functions allowing increased rates of both mutation and total gene loss.

Response: As described, accelerated evolution in CPR/DPANN is not directly caused by gene loss/acquisition events; rather, these genomic changes are downstream consequences of their host-dependent lifestyle, which relaxes selective constraints on core functions. Now the sentence has been restructured to explicitly state this mechanistic hierarchy: "DPANN and CPR often evolve faster than free-living microorganisms, as their host-dependent lifestyle may lead to elevated mutation rates and loss of core functional gene."

Ln 450 – 451: This needs rephrasing, DPANN are not closely related to the Asgard archaea from which Eukaryotes emerged and the authors have not

shown a direct evolutionary link between the viruses of DPANN and those of Eukaryotes. This statement should be modified to reflect that the author's data suggest acquisition of these characteristics by viruses may predate the divergence of Eukaryotes from Archaea but that there is not a direct evolutionary link between the viruses presented in this study and eukaryotic viruses with these characteristics.

Response: We acknowledge that our data do not demonstrate direct evolutionary descent between DPANN viruses and eukaryotic viruses. The sentence has been restructured to frame viral trait similarity as evidence for ancient origins or convergent evolution: "While this may indicate deep evolutionary roots potentially antedating eukaryote-archaeal divergence, convergent evolution cannot be ruled out."

Ln 457 – 459: I see no evidence in the reference provided that CPR genomes are relatively more streamlined than DPANN genomes nor do I see that the original authors made this claim. Could the authors clarify what they are basing this claim on?

Response: We appreciate the reviewer's careful scrutiny. Although Castelle et al. (2018) do not explicitly state that CPR genomes are more streamlined than those of DPANN, their observation that *Ca. Micrarchaeota* and *Ca. Parvarchaeota* (key AMD-associated DPANN lineages) possess "more extensive metabolic capabilities" implies greater metabolic complexity in these DPANN groups. To quantitatively assess streamlining, we estimated genome sizes using data from Castelle et al.'s Supplementary Table 4 and the provided formula: $(\text{Assembly Size} - (\text{Assembly Size} \times \text{Contamination\%})) / \text{Completeness\%}$. Comparative analysis revealed significantly smaller median genome sizes in CPR (Supplementary Fig. 22). Our own comparison of genome sizes in AMD sediments, along with METABOLIC-based functional gene analysis showing fewer genes in CPR, further supports the conclusion that CPR genomes are more streamlined in this system. To avoid ambiguity, we have limited this

discussion to AMD sediments (lines 477–478).

Ln 515-516: This sentence is unclear, are the authors saying that if the viral genome is aligned to the host genome then it is classified as a prophage?

Response: The sentence now reads (Lin 548-549): "Prophages are identified as viral genomes that are precisely integrated within the scaffold of the host genome."