Dear Pr. Sébastien Massart,

Thank you for your and two reviewers comments on our manuscript PCIInfections #77 entitled "**African army ants at the forefront of virome surveillance in a remote tropical forest** ". We appreciate greatly that all reviewers and editors found the study to be of interest, as we similarly appreciate the criticism and guidance that the reviewers and editors have provided. We are now submitting a new version in which the text of the manuscript has been modified according to the comments provided, as we explain below point by point.

Managing Board of PCIInfections comments:

1) **Data, script and code availability (if applicable)**

We have now deposited scripts and codes that were used in this study in Zenodo with a DOI.

2) **Supplementary information (if applicable)**

Supplementary information (tables and figures) are now also deposited in Zenodo with a DOI.

Recommender comments

1. Introduction

- 87 millions of eukaryotic virus species on earth

This number is an estimation of the number of viral species per eukaryotic species multiplied by the estimation of eukaryotic species on earth. This double estimation has a very large uncertainty and I suggest to eliminate the number but simply indicating there are millions of viruses which already illustrates the gaps with ICTV recognized species

We agree with Recommender's comment. We have eliminated the number "87".

- The organisms that we farm

Thamed organisms, being animals or plants could better represent the borders as dogs or cats are not farmed for example

We have reworded the sentence as follows: "… the organisms that we tame and farm."

2. Material and methods

- Positive and negative controls have been used, which is very positive and can rise the confidence in the obtained results but what were the positive and negative controls used?

Five negative controls containing each 8 ml of 1x Hanks' buffered salt solution were added to the three libraries. We are sorry to note that we have made a mistake in the first version of the manuscript because only negative (and not positive) controls were

added to the libraries. We have performed read analyses of the five negative controls and the number of virus reads were determined. Indicative of cross-sample contamination, virus reads assigned to the virus families mostly represented in the ant samples (i.e. *Parvoviridae*, *Microviridae*, *Dicistroviridae*, *Cruciviridae*, *Circoviridae*, *Iflaviridae*, *Polycipiviridae*, *Retroviridae*, *Bidnaviridae* and *Nodaviridae*) were found associated with the negative control, with a mean of 15 reads per virus family. This result suggests that a minimum of 15 reads assigned to these 10 virus families would be a conservative threshold above which a sample should be considered as likely containing viral sequences assigned to these 10 families. On the other hand, no "read threshold" was used for the virus families for which no evidence of cross-sample contamination was identified. We have modified the revised version of the manuscript including the read analyses of the negative controls.

- The minimal length of the contig is well justified but the e-value threshold not. Can it be explained ? Indeed, there might be a risk of detecting Enogenous Viral Elements (EVEs) from genome sequences of the hosts with such value.

We have used a 10-3 E-value threshold, which has generally been widely used in VANA studies. However, other more stringent criteria have also been used (e.g. an E-value threshold of 10-4 in Ma et al. 2019. Journal of virology 94:e01462-19). We can not totally rule out that EVEs were not inventoried using this e-value threshold but while DNAse were used during the semi-purification of the virus particles, the majority of host DNA should have been removed, including host DNA regions containing EVEs.

- How can the link be done between the supplementary table 1 (identifying each sample) and the raw data presented in SRA, more specifically the internal tags identifying each sample within a library (e.g. the 3 pooled sequencing dataset MGN-1, MGN-2 and MGN-3 by Illumina and Flongle sequencing)) ? I could not find it. So adding a column in Supplementary table 1 with the corresponding tags used would facilitate reanalysis of the data of this pioneering sequencing effort by the scientific community

We thank the Recommender for having highlighted this point. We have modified the Supplementary Table 1 and we have now added a column with the corresponding SRA accession number. While "cleaned" reads (tags and adapters removed from the raw reads) were deposited at the ncbi SRA website, PCR Tags that were used can no longer be found in this dataset. However, mapping of clean reads on the contigs deposited in GenBank can be more easily achievable using the "clean reads" dataset.

3. Results

- There is no information on the results of the controls and how it helped in results interpretation (as it has been considered as the third step of bioinformatic analysis in a recent publication – DOI: 10.24072/pcjournal.181 - and in a new EPPO standard PM7/151.

As mentioned above, we have now clarified what type of analyses of the negative controls we have done.

- This point is related to reviewer 1 comment concerning the ~24,000 contigs potentially of viral origin but without similarity to viral genera recognized by ICTV: what to do with them ? They are not really discussed while they could have a great interest in filling the knowledge

gaps between existing viruses on earth and already discovered ones (although I acknowledge it is important to remain cautious about them and not being too speculative).

Among the 46,377 contigs that exhibited sequence similarity to viruses (BLASTx e-values < 0.001), 11,146 contigs were assigned at the viral realm level, 1,377 were assigned at the viral order level and 11,448 were assigned at the viral family level. Finally, 22,406 of the 46,377 contigs (48.3%) exhibited sequence similarity to viral genera recognized or in the process of being recognized by the International Committee on Taxonomy of Viruses. We have modified the text of the revised version of the manuscript and this information is now included.

- Were PeVD and SoMV the only known plant viruses detected ?

Contigs with detectable homology with plant virus families containing economically relevant crop pathogens (e.g., *Reoviridae*, *Tombusviridae*, *Geminiviridae*, *Solemoviridae*, or *Alphaflexiviridae* Figure 1) were also identified. We have reworded the paragraph dealing with plant viruses and we present now that PeVD and SoMV were only two examples of putative detected plant viruses that we decided to focus on.

- While using VANA, how do you explain the small contigs retrieved from plant viruses ? It means complete viral particles were not recovered or the sequencing depth was not high enough as the initial concentrations of the plant viruses were too low ? This could be discussed (maybe more broadly for viruses infecting non-arthropod hosts)

We thank the Recommender for these questions, we have slightly more discussed this point in the revised version of the manuscript focusing on potentially the insufficient sequencing depth of low abundance viruses within complex viral communities composed sometimes of more than ten viruses, some of which being highly abundant.

Reviews

Reviewed by Mart Krupovic, 03 Feb 2023 17:28

In this manuscript, Fritz and colleagues explore the virome of army ants from a tropical forests. The overarching objective of the study was to test the hypothesis that army ants, obligate collective foragers and group predators, can be used as proxies for random/unbiased virus sampling in difficult to access areas, such as densely forested tropical regions. Using 209 army ant samples collected from 29 colonies the authors have discovered a staggering number of different virus species from 157 genera in 56 viral families, some of which are predicted to infect the ants whereas others are linked to various food sources. Thus, this work clearly shows that ‘proxy sampling’ using army ants or other highly mobile predator/scavenger animals is a viable approach which can provide valuable information on virus diversity and ecology. The manuscript is very clearly written and I enjoyed reading it. I have a few questions/comments which the authors could consider.

All viruses which are described fall into existing taxa (above species/genera). I am curious whether the authors have identified something “new(er)”. Please comment: is this information is retained for subsequent publications or there are some biases in analysis/sequencing which

precluded identification of novel virus groups or have we already sampled this part of the virosphere deeply enough?

As mentioned above, we have found 46,377 contigs that exhibited sequence similarity to viruses (BLASTx e-values < 0.001), among which 11,146 contigs were assigned at the viral realm level, 1,377 were assigned at the viral order level and 11,448 were assigned at the viral family level. If we quote what Stobbe and Roossinck have proposed years ago: "The viruses found using metagenomic sequencing data can be described in three different ways: (1) Known-knowns: virus species or isolates that are already known to be in the environment being surveyed; (2) Unknown-knowns: new virus species or isolates of a known family, or known viruses that have not been found previously in the surveyed environment and; (3) Unknown-unknowns: viruses that are completely novel and share little to no sequence similarity with other known viruses.", then we have investigated in this study Known-knowns and Unknown-knowns viruses but we have put aside Unknown-unknowns viruses that are potentially hidden in the approximately 24,000 sequences that were assigned at the realm, order and family levels. This information is not retained for subsequent publications but we completely agree with Mart Krupovic that it would be definitively worth analyzing with more details these sequences and see whether novel virus groups would emerge from this original dataset.

Given that some (most ?) viruses are derived from the food source, have the authors considered that there might be a bias towards viruses more resilient to the harsh environment of the digestive tract?

This is a relevant point that we now discuss in the paragraph "Genetic and morphological factors influence the army ant virome" of the revised version of the manuscript. Indeed, we agree with Reviewer#1 that the VANA metagenomics approach, that is based on the analysis of virion-associated nucleic acids, has potentially biased our results towards the semi-purification of viral capsid that were more resilient to the harsh environment of the digestive tract. Consequently, using this approach may have limited the detection of capsidless viruses, such as several mycoviruses.

A related question, can the authors assess the extent of damage/fragmentation for genuine ant viruses versus those coming from the food?

We guess that using the VANA approach can not allow assessing whether viruses were genuine ant viruses or viruses coming from the food. In further studies, combining the VANA approach and the double-strand RNA (dsRNA) approach developed by either Roossinck's group or Candresse's group would allow addressing this question while the dsRNA approach is based on the isolation of dsRNA, a hallmark of RNA virus infection.

Page 10, L12: was there a reason to use neighbor joining for some proteins and maximum likelihood for the others? If so, this could be noted in the methods section.

We initially used the maximum likelihood method for studying the phylogenetic relationships of ant cytochrome oxidase, NS1 proteins of chaphamaparvoviruses and Rep protein sequences of cycloviruses. We then started studying the phylogenetic relationships of the remaining virus families and we noted that the topologies of neighbor joining and maximum likelihood trees were totally similar for microviruses

and parvoviruses. We then decided to use the neighbor joining method which was much faster for the remaining virus families.

Page 10, L13-14: coat protein and capsid protein are synonymous. Why use both terms?

This is right: capsid protein is now replaced by coat protein in the revised version of the manuscript.

P17, L4-5: ""unclassified Caudoviricetes" subfamilies (recently abolished Myoviridae, Podoviridae and Siphoviridae families)" – this is confusing. The abolished myo, podo and sipho did not become subfamilies.

We have removed the word subfamilies and keep the term "unclassified Caudoviricetes" as recommended by Turner et al. Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. Archives of Virology (2023) 168:74

P17, L18: change "that is comprised of" to "that comprises".

Done

P17: I understand why picobirnaviruses are listed under "Bacteriophages" subheading, but it is not clear why cruciviruses are also listed in this section. As far as I know, there is no evidence suggesting that these viruses infect prokaryotes.

This is true. We have now created a new section named "stramenopiles/alveolates/Rhizaria viruses" where cruciviruses are now listed.

P19, L8: What was the mean size of other contigs?

We have added in the revised version of the manuscript the mean size of the contigs assigned to bacteriophage families, to the *Cruciviridae* family, to invertebrate virus families and to vertebrate virus families.

P20, L4-7: From the way it is written, it seems that the authors suggest that the detected plant viruses are derived from the eaten herbivorous insects. Is this what was meant? Can the authors exclude the possibility that they come directly from the plants consumed by the ants?

No, the plant viruses may have come directly from the plants consumed by the ants. We have modified the sentence in the revised manuscript.

P20, L21: Perhaps delete "new". It has been "new" for more than a decade already.

Done

P23, L22: "…SF3 sequences (see for instance the clade located between the two ambidensovirus groups, Figure 4) and…" – please indicate the clade somehow in the figure (arrow? some symbol?).

We have added a symbol (an ant and a question mark) nearby the clade located between the two ambidensovirus groups in the Figure 4.

P27, L15: the authors state that nine new species could be created, but only 8 are listed in the parentheses.

Absolutely! Thanks! This is a typing mistake; we have forgotten to mention Army ant associated cyclovirus 9. Done.

Reviewed by anonymous reviewer, 26 Jan 2023 17:18

Basic reporting

This paper provides an insightful analysis of a virome surveillance. It offers an original approach that is both rigorous and accessible. The findings are mostly well-supported, and it is likely to be highly cited. The paper is an elegant example of research that will be widely recognized.

This article presents an original approach to the virome of a remote ecosystem by using ants as a proxy. With just 209 ants, the authors were able to detect 22,406 virus-like contigs belonging to 56 families. Seventeen of the 29 ant colonies were identified thanks to the "accidental/non targeted" recovery of the COI gene. This approach is likely to lead to an increased level of detection in poorly studied areas. This will be beneficial for the global virome description. Notably, the authors highlighted the overrepresented families of Parvoviridae and Circoviridae. Sequences of 403 Parvovirus were analysed based on their SF3 proteins, with more than 200 amino acids available for comparison with publicly available data. This revealed an increased diversity, as well as an expanded geographical distribution and potential host range. Additionally, 45 complete genomes of novel cyclovirus were resequenced and compared with publicly available data, providing further insights into this family.

We thank Reviewer#2 for his/her nice appreciation of our work.

Experimental design

The work on the Parvoviridae and Circoviridae families is very thourough; however, due to the nature of the sequences, similar conclusions could not be made for other virus groups. The number of contigs and virus families identified in the study were based on contigs with lengths ≥200 nt, and retained viral BLASTx assignations of these contigs with e-values < 0.001 (M&M). It is not specified in the text that the BlastX was done on a complete non-redundant protein database (GenBank non-redundant database is indicated on the legend of fig 1).

BlastX was done against the GenBank non-redundant protein database. This is now mentioned in the revised version of the ms.

The amino acid identity recovered, as reported in Figure 1, was as low as <25%. Figure 1 is informative but can be misleading as a virus species can be represented multiple times, e.g. the two closely related points for the nepovirus can represent two different viruses or two contigs covering different parts of this segmented virus. In addition, the percentage of

homology represented in figure 1 can be from very conserved genes (e.g. RdRP) or from putative genes with low homology even within well described families (the same virus could have multiple contigs with very varied homology to the closest sequence from the database). The legend of this figure should also be clarified as to whether the amino acids homology is per sequence alignment, or the homology given by BlastX, where only the matching region of the molecule is measured (in which case, this can be a fraction of a short 200 nt contigs (67 aa)).

We understand that Figure 1 can be misleading as a virus species can be represented multiple times but we guess that our legend of Figure 1 was rather explicit regarding both concerns of Reviewer#2 (representation of assembled sequence contig and amino acid homology based on BlastX results). Specifically, the legend of the figure was: "Each dot represents an assembled sequence contig with the corresponding protein identity (best BLASTx e-values < 0.001) to plant (top) and animal virus (bottom) in the GenBank non redundant database".

In the manuscript, the authors have been cautious not to overstate their findings. It is evident that ants are a good proxy to access difficult regions, and the authors note that the ants are "not completely unbiased". Judging by Figure 1, they are clearly biased towards animal, mostly invertebrate ssDNA viruses (as mentioned p14L12). Few plant viruses are detected and mycoviruses are not discussed at all. The fact that these viruses have to pass through additional steps in the trophic chain is discussed on page 19, but what can be said about viruses with low stability, concentration, or prevalence? The principle of VANA should yield nucleic acids protected by a capsid (in contradiction with the degradation observed).

As mentioned above, this is a relevant point that is now discussed in the paragraph "Genetic and morphological factors influence the army ant virome" of the revised version of the manuscript. We now say that the VANA approach may have limited the detection of capsidless viruses and viruses with low stability, such as several mycoviruses. In addition, we also pinpoint now that the VANA approach may have hindered the detection of viruses with low concentration or prevalence. Hence, an insufficient sequencing depth of low abundance viruses within complex viral communities composed sometimes of more than ten viruses, some of which being highly abundant may have biased the inventory towards the more abundant/prevalent viruses.

Are ants the best candidates for a plant metavirome? The authors should provide a more detailed discussion about this.

Ants, or at least Army ants, are probably not the best candidates for inventorying the plant metavirome. We now provide a more detailed discussion in the revised version of the ms. at the end of the section "Plant viruses".

While the identification of the ants through the recovered reads matching the COI is a useful bonus, it is not definitive. The number of reads is small, and the VANA tool is not designed to recover non-encapsidated viruses. Additionally, if this experiment was to be repeated, it would be beneficial to have some morphological identification and/or a proper DNA barcoding on the ants (which would require collecting two samples for each species, one for the metagenome and one for the taxonomy).

We fully agree with Reviewer#2's comment. We also thank Reviewer#2 for his/her advices for improving our further studies.

Validity of the findings

It is clear that besides the Parvoviridae and Circoviridae families, the contigs extracted were mostly short and from different genomic regions. In some cases, this allowed for a taxonomic assignment, and presumably, in other cases, the contigs could only be used to make the Figure 1 (137 sequences were deposited on GenBank for the phylogenetic analyses out of the 22,406 contigs). But that is the nature of these metagenome studies. Therefore, I understand that the phylogeny used is there to illustrate that the virus contigs (or virus-like in some cases) fit into available taxonomies but it would be good to explain why neighbor joining method was chosen.

Please see above our answer to this point already raised by Reviewer#1.

Additional comments

There are a few additional small edits:

Page 3 L23: The sentence needs to be rewritten as it reads as if the viruses have medical or agricultural relevance to human (instead of the host).

We have reworded the sentence: "Hence, samples were only accessible via forest roads or tracks, and derived from the subset of plant or animal species that are likely to host viruses with some medical or agricultural relevance".

Page 4 L14: Densely forested tropical regions do not represent major interfaces but rather provides interface as a consequence of human activities surrounding these forests. Additionally, densely forested tropical regions clustered together, represent fewer interfaces than if the forests were scattered across a larger territory.

We have also reworded this sentence.

Page 4 L21: random/unbiased : all the tools relying on one animal will have preferred patterned, but those will be different to the human one. I like the way it is defined earlier "a less human-centric assessment of viral diversity at the ecosystem-scale"

We have also reworded this sentence and replace "random/unbiased sampling" by "human-centric assessment of viral diversity".

Page 25 L4: Since endogenous paroviral elements are detected within invertebrate genomes, how many of the parvovirus contigs could be EPVs?

As mentioned above, we can not totally rule out that EPVs were not detected in our VANA study. However, while DNAse were used during the semi-purification of the virus particles, the majority of host DNA should have been removed, including host DNA regions containing EPVs.

Figures with phylogenetic analyses (mostly 2 and 3): could you when the aligment is made on the protein or the nucleic acid and the size of the region aligned.

We have included this information in the legend of Figures 2 and 3.