Comments from the recommender:

We thank the recommender for the very careful reading and helpful suggestions on the manuscript.

Summary

- In line 30 there is a reference to xGen kit which is a commercial kit. Could you instead explain the biochemical mechanism on which the kit is relying ?

1. Done (now line 32 and everywhere else)

- In line 35: the S before MDA should be explained (it is only defined in line 67)

2. We replaced here s by short, line 37.

- In Line 36: is the direct DNA shearing linked to xGen kit or to the original shotgun sequencing ? It should be clearer (it is understood further in the document but the summary is self standing).

3. DNA shearing is a step of the xGen kit, it is now rephrased to make it more understandable (Line 32).

- Line 37: “including two with ssDNA” adapted to “among which 3 with dsDNA and 2 with ssDNA genome” ?

4. Rephrased and clarified (line 40)

- Results: it is mainly material and methods while the results are very (too ?) synthetic. Please adapt the text.

5. Exact. A Methods section is now introduced in the summary, line 36.

Introduction

- Line 59: low amount of DNA or of viral DNA ? Please clarify

6. It is viral DNA, now clarified, line 62.

- Line 69-71: same point as abstract: explaining biochemical/technological principle and referring to commercial kit further on (inverting order with following lines explaining the protocol)

7. Done, lines 73-78.

- Lines 76-79: is there any scientific reference for it ?

8. Our reasoning was based a common scientific assumption: more energy is needed to break two strands versus one strand, and it proved to be correct.

- Lines 82-86: these are already results. Could you instead summarize the methodology (testing pure virome + virome with cells) ? + comment for lines 92-93

9. The final paragraph of the introduction is now changed to fit these recommendations, lines 86-92.

Material and methods

- Lines 92-93: objective statement would fit well at the end of the introduction instead of methods.

10. Done, the first paragraph of the Methods is suppressed.

- Lines 92-93: The objective isn’t it for comparing protocols of ssDNA enrichment (the mixes are the tools to reach the objective)?

11. Yes, rephrased and put at the end of the Introduction, ines 90-92.

- Lines 97-98: can the third method be considered as a control?

12. It is not a real negative control, as this kit is designed to recover ssDNA fragments.

- Line 101: abbreviate NEB at first occurrence.

13. Done

- Line 109: is there a reference (publication, official repository) for strain C?

14. Yes, added, line 110.

- Lines 107-110: the two phages should be presented in the first sentence (that only describes PhiX174) and developed in two independent sentences further on

- Lines 111-116: same as for ssDNA: citing the 3 phages in first sentence and describing each of them next.

- Line 119: citing the provider of PES membranes.

- Line 120: reference to LB medium (provider).

15. All improvements added.

- Line 121: is there a temperature during centrifugation? or room temperature?

16. Yes, 4°C, this is now added.

- Line 132: what is SM buffer?

17. A very old buffer name, it means salt-magnesium. Now explained, line 123.

- Line 133: $CaCl_2$ not correct form for 2

- Lines 136-137: centrifugation steps not described for PCA extraction.

18. All done.

- Phage stock preparation and DNA Extraction:

o overall, it is not clear for me what has been carried out on which phage. Indeed, there is the description of a DNAse treatment in lines122-125 and another one in lines 131-133 (including a RNAse).

o Phage DNA preparation only related to T4, SPP1 and lambda (Line 128): what about PhiX174 and M13-ypf ?

19. This whole section of the Methods has been revisited, and sub-divided to help the reader follow all the steps involved.

- Line 139: estimating a concentration using nanodrop (for ssDNA) is not so reliable. Could you provide evidence that nanodrop allowed accurate concentration evaluation? Why not using another system such as Qubit ssDNA Assay Kit? (in link with comments of a reviewer)

20. Theory predicts that a UV-based method is more appropriate, for ssDNA concentration estimates, than a method based on intercalating agents. In our knowledge, the problems of Nanodrop measurements are (i) over-estimates due to phenol or other compounds remaining in the DNA sample, (ii) wrong estimates for genomic and viscous DNA. We never noted problems with ssDNA. It should be noted moreover that the ssDNA of PhiX174 and M13 was purchased from NEB, and we just verified its concentration with the Nanodrop.

Since two of the referees stumble on this question, we ordered and tested the ssDNA QuBit kit, and obtained the following measures:

	Nanodrop		QuBit	
	Measure 1	measure 2	Measure 1	measure 2
ssDNA PhiX174 NEB, 100 ng/µL	101.6	95	76.2	67.6
ssDNA M13mp18 NEB, 25 ng/µL	18.8	18.5	15.7	15.4
Standard ssDNA of the QuBit kit, 20 ng/µL	23.8	23.1		

We conclude Nanodrop measures are closer to the theoretical concentrations (taken from the10-fold dilution of NEB DNA concentrations, as stated on the tube). Nanodrop repeats are less reproducible than the QuBit ones. We also note there is a 15% overestimation of the Qubit control tube by the Nanodrop apparatus. Taking this into account, Qubit and Nanodrop values are congruent for M13, but PhiX174 continues to be some 20% underestimated by

QuBit. In summary, we think we were on the safe side with our Nanodrop measures of ssDNA concentration.

- Lines 141-142: is the protocol identical to the one described just above or not?

21. Yes, this section is part of what has been reformulated (see point 19).

- Lines 143-144: why using different volumes for each treatment ? It could introduce bias potentially. Can the authors discuss it (explaining also the rationale of the choice - potentially linked to manufacturer instruction?)

22. Thank you for this question. There was a mistake, 19µL should read 5µL. The 1µL volume used for sMDA is indeed a manufacturer’s instruction. To help the reader follow the procedure, we now mention what were the final DNA yields, lines 172-174.

- Line 148 and elsewhere: stating supplementary table 1, 2.. in full letters

- Line 155: provider of the kit (overall comment: please double check that the providers are mentioned when using kit/enzyme/reagent)

23. Both done

- Line 192: the final temperature was always 46°C after 10 minutes at room temperature?

24. The important point here is to reach a temperature below 46°C, to ensure proper annealing of primers. We have now rephrased this, line 186.

- Line 203: consistency when citing the provider of phages and DNA (adding reference number as before)

25. Most references of this type are now given just once, in the first Methods paragraph.

- Line 215: a mix of 6 PCR fragment is mentioned but its origin is not clear as there is no PCR step mentioned in this paragraph (if produced elsewhere, please indicate how or a reference or their size …)

26. We now mention their sizes, lines 212-213. These fragments were “in house”- available PCR fragments that we collected for the experiment.

- Line 228-232: there is no indication of mix or reference to publication for the details on the methodology and reagents used for xGen kit (beyond temperatures and time)

27. We followed the manufacturer’s instructions (this is now added line 226).

- Line 234-235: this depth also concerns the fecal samples? Better to move this depth in results, integrating fecal samples.

28. No the virome samples were sequenced at a different depth of 5 million reads, we now specify the sequencing depth for each set of samples lines 232 and 262.

- Line 251: is there any specific buffer for this treatment? If so, please mention it

29. No, benzonase is an enzyme that does not need buffer and works in very crude solutions.

- Line 264-264: please clarify the two approaches because the software approach also uses databases (Vibrant uses KEGG, pfam and VOG)

- Line 267: a bracket is missing after 21

- Line 273: “of both approaches”

30. All done.

- Lines 273-275: it is another approach used (5th one), is it done downstream of another approach or it also starts from contigs. ?

31. We now explained more lines 283 to 290. To detect *Microviridae*, we start from the set of already detected viral contigs (so it is not an additional, 5th approach to detect viral contigs!), we then perform gene calling on them with Prodigal, and we finally compare all these proteins to the three profiles of proteins typical of Microviridae, using HHpred, and retaining hits with a probability above 90% and a coverage of the target protein above 60% (conditions used in the PHROGS database).

- Bioinformatics analyses of fecal samples :

o a flowchart indicating the analyses done could help having a global picture (as further mapping of normalized reads was carried out). It is not clear how to reproduce the succession of analyses and it needs clarification.

32. OK, done and placed as supplementary figure 1.

o There is a set of viral contigs for mapping but how is it selected as several methods are used to identify the viral contigs: any viral contig from any method OR only viral species identified by all methods OR …. Please clarify how the set of contigs has been built up.

33. We have now clarified line 277 that we retained any viral contig from any predicting tool (with the mentioned parameters).

- Lines 276: all samples -> “The four samples”.

34. Done

Results

- Line 304: nanodrop is not really reliable for accurate quantification (same comment as before and as a reviewer)

35. See answer number 20.

- Line 325: I do not find the mix’ explanations (1:1 in volume? in quantity?) in Methods

36. It was in the T7pol treatment section, in the Methods. It is also mentioned now in the result section, line 341.

- Figure 2: is the standard from a commercial provider or made internally ?

37. It is a commercial one, now specified in the legend of figure 2.

- Setting up a ssDNA-to-dsDNA conversion protocol using T7 DNA polymerase: I do not see any replication of the test nor any test with other organisms (or other proportion between both viruses). There is therefore no information on the reproducibility of the observations. Next chapter shows it worked but can the authors explain why there is no replication carried out (for ensuring the selection of 25 µM for example).

38. Thank you for raising the point. There was a progressive setting up of the conditions found more appropriate for T7pol-dependent replication from a mix of degenerated primers. We mostly investigated the question of the primer concentrations, and did not vary the proportion between the viruses, nor tried other genomes.

The exact combination of concentrations shown in Figure 2 had not been repeated 3 times on the same gel, so we have done it now, and a gel with a better separation of the products is shown in place of the previous one. We also noticed a mistake in the computation of the range of concentrations: they were 2.5-fold lower than indicated overall in the manuscript. The optimal concentration of oligonucleotides was 10µM, not 25µM as stated erroneously. It is very high, and we understand it can raise doubts. However, a theoretical calculation of the amount of 20-mers complementary to M13 present in the 10µM reaction indicates that even at this high concentration of degenerated primers, only a minor fraction (10^5 primers, for a total of $6x10^{10}$ M13 molecules present in 250 ng) of the ssDNA could be converted. It suggests that priming takes place also on partially annealed oligonucleotides. For instance, if one assumes that a good annealing of the last 10 nt of the oligonucleotide is sufficient for priming, then one expects 10^{11} primers available for M13, to convert the $6x10^{10}$ molecules of the reaction.

- Line 334: “way of treating DNA” -> protocols

- Line 338: see Suppl. Table 3 enough between brackets

- Figure 3: the legend can be completed referring to Rel value. A distinction in wording between the theorical initial proportion (see comment on quantification protocol) and observed proportion after high throughput sequencing is welcome.

39. All three suggestions done.

- Figure 3: “Various” sample is vague.

40. Reformulated more precisely.

- Suppl. Table 3: distinguish ssDNA from dsDNA phages to facilitate reading of the table.

41. Done

- Lines 349-358: this is mainly a repetition of the methods. It is very clear so it can be fused with duplicated information in methods.

42. We wanted the reader to understand the process in its great lines, without needing to refer to the Methods section. We now shortened this section, while still giving an outline.

- Line 376: Aitchison distance calculation is not described in Methods while used in results.

43. A paragraph is now added at the end of the "*Library preparations, synthetic phage mix sequencing and analysis*" section, in the Methods, lines 246-250.

- Line 390: "treated or not with…"

- Line 393: "assembled in…"

44. The two changes were done

- Viromes analyses: globally, numbers would be interesting (number of reads, of contigs, of viral contigs assigned by each algorithm or in total).

45. This human virome analysis is part of a project with a larger number of samples, which will be presented later, in a distinct manuscript. This made it difficult to present the numbers all along the pipeline. We give the final number of viral OTU per sample in the last section of the Methods dealing with them, lines 300-304.

- Line 394: it is not clear how the contigs were assigned to microviridae (among the various pipelines used).

46. See answer 31. We think it is now more clearly stated, in particular in the Methods section, lines 283-290.

- Line 396: *Microviridae* abundance of 24 to 74%: of the total number of reads or of the viral reads ? Stating it has been obtained after mapping is relevant

47. All explained more precisely, lines 407 and 413, thanks.

- Lines 397-399: harmonize the wording: 6 microviridae contigs & 11 microviridae.

48. Done.

- Overall: why the analysis focused only on microviridae and not dsDNA phages ? Indeed, the results on phage mixes show that the reads from dsDNA can drop with T7 polymerase (Lambda in Panel A-2 for example). This could give a global overview of the performance of T7 polymerase treatment

49. *Microviridae* and dsDNA contigs (ie Caudoviricetes) are the two main contributors to human viromes, so that the change in dsDNA contigs relative abundance usually mirrors the *Microviridae* changes. We now added a supplementary figure 3 (and a suppl. Table 5 with the corresponding raw data) with the Caudoviricetes results. As expected, the Caudoviricetes are

under-represented upon T7pol treatment in sample S4, as this sample contains a majority of *Microviridae*, once treated with T7pol. No such trend is visible for sample S18, where *Microviridae* concentrations do not move as much. We now add a sentence to comment on that. We comment on that lines 416-421.

Discussion

- Line 432-433: This is completely true. Nevertheless, there was no replication for some steps of the optimization (for example selecting the primer concentration at 25 µM).

50. See answer 38.

- Line 435: should limit the bias as they still exist based on the presented results

51. Correct, rephrased.

- Line 445: could the 2-fold observation be caused by a saturation in ssDNA in the mix (98%), limiting therefore the further amplification? Does it worth discussing this value?

52. We agree that there is saturation here, and this amplification up to 98%, prevents detecting the dsDNA phages that are present. We now formulate this more precisely, line 451.

- Lines 463-465: this comparison is very interesting and raise the question on the actual potential results of the T7 treatment when using Truseq or Nextera. It should be noted that there is no data supporting it. Could the results be different with these kits or guaranteed similar?

53. We also answer on this point to referee 1 (answer # 56). Indeed, we have no proof for this. The sentence is now tuned down to express a speculation, line 470. The Truseq libraries, which require a sonication step, should lead to results similar to those presented here. The Nextera libraries are based on enzymatic cleavage of the DNA, it may lead to some differences.

- Lines 474-476: this is a very interesting preliminary comparison and linked to the the number of species detected in both samples. Are they comparable to literature or not for microviridae in fecal samples (somehow, 8 species might be very low ?)

54. As nobody knows how to confidently sequence both ssDNA and dsDNA viruses in a virome mix, the litterature is not informative on this point. Our work hopefuly paves the way to a better assessment of this question.

- Line 480: “can improve the sequencing of ssDNA viruses”

55. Yes, much better formulated. Thanks.

by ***Sebastien Massart****, 20 Feb 2024 14:32*
Manuscript: **https://doi.org/10.1101/2022.12.12.520144**
version: 2

Review by anonymous reviewer 1, 06 Feb 2024 06:43

This study aimed at providing an alternative method of environmental VLP-derived DNA preparation before NGS-based virome sequencing. The key point of current suggestion is conversion of ssDNA genome into dsDNA genome using T7 polymerase in DNA preparation step. The authors compared the new method with the previous two methods, MDA and xGen kit, and found that the new method minimized the deviations at least from over-estimation of ssDNA viral genomes by MDA and from under-estimation of ssDNA viral genomes by DNA shearing of xGen kit. The T7 pol method is quite convincing, and is expected that more accurate abundance of dsDNA and ssDNA viruses could be estimated in metagenome studies of environmental virome. The following questions make the manuscript strengthen scientifically more, I believe.

1. A synthetic viral mixture was prepared, and three methods were applied to the mixture to be compared (only xGen kit vs. xGen kit + MDD vs. xGen kit + T7 DNA pol). It is thought that T7 DNA pol with no xGen kit would be considered a standard control in this study, as described "it can be applied to samples planned for any kind of downstream library preparation kit for low DNA amounts, such as the Nextera or TruSeq DNA nano kits from Illumina.". However, the authors set xGen kit as a standard control. I do not think that xGen kit is necessary for T7 polymerase-used virome sequencing, and thus, two groups (no xGen kit, no xGen kit + T7 DNA pol) need to be compared additionally with three methods.

56. We of course understand the interest of exploring further other combinations of the T7pol treatment with various library kits. However, we think this extends beyond the scope of this report, which was aiming at testing whether a dsDNA conversion step with the T7 DNA polymerase, which has no strand displacement activity, contrary to the one of Phi29, would prove advantageous in terms of quantitative sequencing. And we think we have made the point. The matter of exploring other library kits was raised in the Discussion section, to illustrate how others might find it advantageous as well with their preferred library preparation protocol (or with the companies with which they subcontract for library preparations). We really think the pretreatment should be compatible with all possible kits, and rephrased it so as to be clear this is simply a prediction.

2. The Qubit and Nanodrop devices were used for estimating absolute concentration of dsDNA and ssDNA phage stocks. It is quite not sure that measuring DNA concentration using Nanodrop is accurate.

57. As explained above (answer 20), we were confident that for ssDNA, Nanodrop would be a better solution than Qubit. We now tested it, and present the results in answer 20. Basically, Nanodrop is fine.

In addition, the number of M13-yfp phage stock was estimated using plaque assay. Three different methods make some predicting exact number of phage genomes, and the deviation from three different methods may have an impact on assessing the abundance of ssDNA viral genomes from three DNA preparation methods.

58. We have chosen for each phage the best way to quantify it. For four of the five phages, qPCR gave superior results, probably because it takes into account uninfectious particles. These particles are important to take into account, as they will be sequenced as well. For M13, it was shown that this particle is one of the most resisting one to thermal treatments (see for instance https://doi.org/10.1016/j.tca.2018.12.010) and indeed we found more elevated titers

by plating than upon qPCR analysis. This explains why this method was chosen. We have now added this reference to strengthen our choice.

3. According to the previous study (Appl Environ Microbiol, 2010, 76(15):5039-5045), it is thought that conversion of dsDNA into ssDNA works with E. coli DNA polymerase I in the presence of random hexamers. In this study, T7 polymerase was chosen for conversion of dsDNA into ssDNA, instead of E. coli DNA polymerase I that is commonly used in molecular biology techniques. Please, explain kindly why T7 polymerase was chosen for the DNA conversion.

59. Thank you for this reference. We do not deny that other DNA polymerases, such as DNA polymerase I, might have been selected. T7 DNA polymerase was chosen because of its high processivity. Only 5 minutes were needed to complete the replication of up to 7249 nt (the size of M13mp18). A mutant of this polymerase was used also in the early years of DNA sequencing (the sequenase). In the Canceill et al. paper that is cited in reference, a paper that inspired this work, T7 pol was compared to the Phi29 polymerase, and shown to stop after one round of replication.

Review by anonymous reviewer 2, 14 Feb 2024 11:13

The manuscript entitled: “Method for preparing virome DNA that allows sequencing of both double-stranded and single-stranded DNA viruses” is a well-written manuscript that further explores an alternative method of DNA preparation for sequencing both ssDNA and dsDNA viruses.

This manuscript is remarkably precise and rigorous, and offers a clear protocol that will be of great benefit to the scientific community working on DNA phages.

So glad to read this. This is also what we think!

I have a major concern concerning the concordance between the title of the pdf file that I reviewed (Method for preparing virome DNA that allows sequencing of both double-stranded and single-stranded DNA viruses) and the title shown on BioRxiv website (T7 DNA polymerase treatment improves quantitative sequencing of both double-stranded and single-stranded DNA viruses). Same problem for the lists of authors that are different between the pdf file and the BioRxiv website. This is an issue that should be fixed.

60. Fixed, with our apologies.

Apart from this issue, I only have two minor concerns:

Summary/Background: the authors state that “Virome shotgun sequencing only reveals the double-stranded DNA (dsDNA) content of a given sample, unless specific treatments are applied”. They may have said that this statement is true when viral genomes are analyzed from “bulk” metagenomes which include both virus particles and microbial cells. Alternatively, semi-purification protocols of virus particle have proved to be efficient (albeit cumbersome) for better detecting ssDNA viruses. The authors may have mentioned in the introduction that these two alternatives (direct shotgun and virus particle enrichment) metagenomics approaches exist and briefly give insights about ssDNA virus yields using both types of approaches.

61. We respectfully disagree here. Regardless of the way the initial sample is prepared, without or with a filtration step to remove most bacteria, the ssDNA viruses won't be sequenced in theory, if nothing is done to ensure the sequencing of the ssDNA fraction. Historically, scientists who prepared viromes always added a step for converting ssDNA to dsDNA, this is why their viromes were richer (and in fact, sometimes excessively rich!) in ssDNA phages. We prefer to maintain the introduction section centered on the virome studies, but we have a dedicated paragraph of the Discussion section where bulk metagenomics studies are considered. We think that at present it is helpless to try giving an overview of what is the "real" ssDNA yield of viromes, or bulk metagenomes, as a consensus methodology is not yet at hand. We hope by this contribution to pave the way towards more quantitative studies.

In the "T7pol treatment of two viromes" paragraph, you noted that 6 of the Microviridae contigs detected in the untreated samples were not present in the T7pol sample, while conversely, 11 Microviridae were found only in the T7pol sample". This result is illustrated by the Suppl. Table 4. I here have missed a number of elements. I would have liked finding in this Table: the length of the contigs, their taxonomic assignment, the %identity shared between these contig and the Microviridae species stored in the International databases for which they matched, etc. Finally, it would have been welcome to find plylogenetic trees (one using the major capsid protein sequences, and another one using the replication gene sequences). This would have helped figure out whether the "missing" contigs clustered in the phylogenetic tree or were scattered around it.

62. The additions suggested for the *Microviridae* are now supplied as supplementary table 4. We could not make the suggested phylogenetic trees for the two proteins, as some of the contigs were missing either one of them. Instead, we did a Viptree (supplementary figure 5) that presents overall the proportion of shared proteins between contigs. One of the 6 missing contigs (microvirus_6 in the T7pol set) clusters with two contigs that are present (microvirus_2 and _3). But the other do not. We think, as suggested in the Discussion, that quantitative virome analyses are hard to achieve, and that this low level of fluctuation in composition is acceptable.

P3L76: ss- and dsDNA

P4L107: I would have written "is a member of the *Microviridae* family" rather than "a Microviridae". This correction would need to be done throughout the ms.

P4L110: *Inoviridae* in italics

P7L248-256: Indicate that the "sample numbers" of the two healthy donors are "S4" and "S18".

63. All this is now corrected, thank you.