Improving the sequencing of single-stranded DNA viruses: Another brick for building Earth's complete virome encyclopedia

Sebastien Massart based on peer reviews by Philippe Roumagnac 🆔 *and 3 anonymous reviewers*

Cite this recommendation as:
Massart, S. (2024) Improving the sequencing of single-stranded DNA viruses: Another brick for building Earth's complete virome encyclopedia. *Peer Community in Genomics*, 100335. 10.24072/pci.genomics.100335

The wide adoption of high-throughput sequencing technologies has uncovered an astonishing diversity of viruses in most biosphere habitats. Among them, single-stranded DNA viruses are prevalent, infecting diverse hosts from all three domains of life (Malathi et al. 2014) with some species being highly pathogenic to animals or plants.

Sequencing of single-stranded DNA viruses requires a specific approach that usually leads to their over-representation compared to double-stranded DNA. The article from Billaud et al. (2024) addresses this challenge. It presents a novel and efficient method for converting single-stranded DNA to double-stranded DNA using T7 DNA polymerase before high-throughput virome sequencing. It compares this new method with the Phi29 polymerase method, demonstrating its advantages in the representation and accuracy of viral DNA content in well-defined synthetic phage mixtures and complex human virome samples from the stool. This T7 DNA polymerase treatment significantly improved the richness and abundance of the Microviridae fraction in their samples, suggesting a more comprehensive representation of viral diversity.

The article presents a compelling case for testing and adopting the T7 DNA polymerase methodology in preparing virome samples for shotgun sequencing. This novel approach, supported by comparative analysis

1

with existing methodologies, represents a valuable contribution to metagenomics for characterizing virome diversity.

References:

Billaud M, Theodorou I, Lamy-Besnier Q, Shah SA, Lecointe F, Sordi LD, Paepe MD, Petit M-A (2024) T7 DNA polymerase treatment improves quantitative sequencing of both double-stranded and single-stranded DNA viruses. bioRxiv, ver. 4 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2022.12.12.520144

Malathi VG, Renuka Devi P. (2019) ssDNA viruses: key players in global virome. Virus disease. 30: 3–12. https://doi.org/10.1007/s13337-019-00519-4

Reviews

Evaluation round #2

DOI or URL of the preprint: https://doi.org/10.1101/2022.12.12.520144
Version of the preprint: 3

Authors' reply, 17 June 2024

Download author's reply
Download tracked changes file

Decision by Sebastien Massart, posted 11 June 2024, validated 11 June 2024

Mansucript greatly improved with minor additional questions

Dear authors,

Thank you very much for resubmitting the new version of your research work.

Two anonymous reviewers and I have evaluated this resubmission. The adaptation improved the quality of the document. There are still some minor clarifications that are welcome before the final recommended version. You will find my few comments included in the attached document.

Another point of utmost importance is the availability of your raw data. I double checked them and it appeared the raw data only include the first experiment with phages while the virome data are not available. The open access of raw data is not only a requirement for high-quality publication, but it also plays a crucial role in advancing scientific knowledge. I kindly ask you to make these data available.

Thanks again for considering PCI in Genomics for your publication.

Kind regards,

Sébastien Massart **Download recommender's annotations**

Reviewed by Philippe Roumagnac, 28 May 2024

Dear Recommender,

My mistake, I do agree with the authors that the ssDNA viruses won't be sequenced in theory, if nothing is done to ensure the sequencing of the ssDNA fraction. Therefore, the authors have fixed all my minor concerns. From my side, this revised version of the manuscript can be accepted.

Best regards,

Philippe

Reviewed by anonymous reviewer 3, 09 June 2024

The authors fulfilled the issues raised by the reviewers including me.

Evaluation round #1

Authors' reply, 19 April 2024

Download author's reply
Download tracked changes file

Decision by Sebastien Massart, posted 20 February 2024, validated 21 February 2024

An interesting protocol evaluation that merits a revision

Dear Authors,

On behalf of the board of Peer Community in Genomics, I would like to thank you for considering it for sending your publication.

Two anonymous reviewers and I have read your document and found it interesting while deserving significant improvements. Indeed, there are currently several comments and suggestions that could improve the manuscript.

You can find the comments and suggestions of both reviewers and myself in this response and we thank you in advance for answering these point by point while adapting the text when necessary.

Kind regards,

Sébastien Massart

Comments from the recommender:

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 ?

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

- 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).

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

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

Introduction

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

- 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)

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

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

Material and methods

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

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

- Lines 97-98: can the third method be considered as a control?
- Line 101: abbreviate NEB at first occurrence.
- Line 109: is there a reference (publication, official repository) for strain C?
- 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).
- Line 121: is there a temperature during centrifugation? or room temperature?
- Line 132: what is SM buffer?
- Line 133: CaCl2 not correct form for 2
- Lines 136-137: centrifugation steps not described for PCA extraction.
- 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 ?

- 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)
- Lines 141-142: is the protocol identical to the one described just above or not?
- 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?)
- 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)
- Line 192: the final temperature was always 46°C after 10 minutes at room temperature?
- Line 203: consistency when citing the provider of phages and DNA (adding reference number as before)
- 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 …)
- 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)
- Line 234-235: this depth also concerns the fecal samples? Better to move this depth in results, integrating fecal samples.
- Line 251: is there any specific buffer for this treatment? If so, please mention it
- 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"
- Lines 273-275: it is another approach used (5th one), is it done downstream of another approach or it also starts from contigs. ?
- 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.

4

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.

-        Lines 276: all samples -> "The four samples".

Results

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

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

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

-        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).

-        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.

-        Figure 3: "Various" sample is vague.

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

-        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.

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

-        Line 390: "treated or not with…"

-        Line 393: "assembled in…"

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

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

-        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

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

-        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

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).

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

-        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?

-        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?

-        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 ?)

-        Line 480: "can improve the sequencing of ssDNA viruses"

Reviewed by anonymous reviewer 1, 06 February 2024

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 dsDNA genome into ss-DNA 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.

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. 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.

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.

Reviewed by anonymous reviewer 2, 14 February 2024

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.

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.

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.

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.

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".