Response to reviewers

Nadège Guiglielmoni, Ramón E. Rivera-Vicéns, Romain Koszul and Jean-François Flot

We thank the recommender and the reviewers for their comments on our article, that we carefully examined and integrated in a new version. We have elaborated some sections, and more specifically the Introduction to clarify the goal of this paper. We also added a new section with recommandations for genome assembly.

Recommender

Dear Dr Guiglielmoni,

I have been through your manuscript, as well as 4 independent reviewers, and we all agree that the manuscript is of high interest.

They all, however, highlighted minor comments before acceptance of the manuscript, that I encourage you to perform quite fastly before I can accept if.

In addition, Dr Bourgeois discussed a lot on different aspects of the manuscript that in my opinion are of great interest. Indeed, proposing specific tools for each step would be of great help for non-specialists and beginners...

However, based on my own experience, such recommendations, while of high quality at the given time of the publication and on some specific genomes, would be quite fastly outdated and may be misleading to readers.

Thus, these comments, while very interesting, are for me to be the subject of an online list that can be quickly updated. I would then propose that you discuss them in the manuscript in this way.

Reviewer 1

First I would like to thanks the authors for the work they have done. Here they present a review paper about sequencing non-vertebrates genomes. As a whole, this paper is very pleasant to read.

Each part is rich of details on history of technologies and methods. Presentation of tools is quite exhaustive. Those two arguments made this paper an excellent starting point for non familiar people with sequencing technologies and more particularly for sequencing non-vertebrates genomes.

In figure 2, I would recommand to use some color to make the message easier to understand, and to use a monospace police for the consensus part.
We modified Figure 2 accordingly.

The central part of the figure 5 might be improved, maybe with clear arrows direction and starting point.
We modified Figure 5 accordingly.

Reviewer 2

Guiglielmoni et al. is an informative and thorough review about the approaches developed in the last years to sequence and assemble genomes with a specific focus on invertebrate genomes. I think both the structure and content (including the numerous suggestions for tools) of this review to be of great relevance and interest for the genomic community that deals with non-model organisms. My comments are very minor, they are actually only suggestions to include some particular points/references or to rephrase small parts of the text. Since there are no line numbers I tried to give clear indications about the text location (as clear as possible).

Page 2, end of 3rd paragraph: The sentence “Many phyla with less direct human implications, however, do not even have a single good-quality genome assembly available to date (e.g., chaetognaths).” may be further supported by Hotaling et al. 2021 that explores (among other things) which phyla are lacking any type of genome assembly https://doi.org/10.1073/pnas.2109019118

We added this reference.

Figure 1: consider to replace "," in the legend with "+" so to make clearer that "Short reads, long reads" means that a combination of those technologies was used to build the assembly.
We modified the legend.

Page 5, end of page: it would be important to highlight that the greater accuracy given by HiFi is obtained at the expense of the length of the reads themselves that must be shorter than the ones used for "regular" PacBio.
It is indeed mentioned in the publication of Wenger et al. thus we completed the text.

Page 6, end of 2nd paragraph: "In addition, secondary metabolites associated to DNA molecules, or branched DNA structures, can also disturb the sequencing reaction." This is an interesting point that I heard discussed many times and it would be nice to gather some references to support it if possible.
We have searched for references, however these negative results are usually not published.

Page 14, 3rd paragraph: "To improve the contiguity of an assembly, contigs can be grouped, ordered and oriented into scaffolds.". I think that the concept of contiguity should be reserved to unfragmented sequences (contigs) and that the process of scaffolding does not really improve the contiguity aspect since scaffolds are, by definition, clusters of sequences bridged by gaps (therefore fragmented). I agree that an assembly with thousands of little separated contigs is much worse than an assembly where these contigs are grouped into a bunch of scaffolds but still the contiguity would remain the same, what changes is the representation of the genome/chromosome structure. I would advise to rephrase the topic sentence of this paragraph to have a less ambiguous meaning of "contiguity". Also, I think that the authors used the equivalence "contiguity = measure of quality/fragmentation = contig N50" in the first part of the review (e.g., Figure 1) so I suggest to keep this one meaning throughout the paper.
Page 17, beginning of the 2nd paragraph: specify which N50, I guess "Contig N50" (?)
We use the terms contiguity and N50 for both contigs and scaffolds, as it has been done in previous publications (for instance, https://doi.org/10.1101/gr.126599.111). It is however important to take into account the number of gaps in scaffolds, in addition to the contiguity (which can be improved with more contiguous contigs and gap filling). However, the definition of N50 only mentioned contigs, and not scaffolds, which was corrected.

Page 15, middle paragraph: regarding linked reads and the discontinuation of the 10X Genomics service, it can be added that there are at least two other replacing technologies: 1) TELL-seq https://genome.cshlp.org/content/30/6/898 ; 2) haplotagging https://www.pnas.org/content/118/25/e2015005118 The second was used already on invertebrates (butterflies) and TELL-seq seems to work with ultra-low DNA input.
Thank you for this suggestion.

Page 16, end of first paragraph: the last sentence of the paragraph could be slightly rephrased in a way that becomes 100% clear that using Omni-C can yield de novo genome assemblies. What I mean is something like this: "[...] such as Omni-C, therefore adequate for de novo genome assemblies." This is just an example, no need to rephrase it exactly like this!
We clarified this sentence.

Reviewer 3

I read the manuscript titled A deep dive into genome assemblies of non-vertebrate animals by Guiglielmoni et al. with great interest. The authors talk about existing methods and algorithms for constructing contiguous and accurate genome assemblies in the context of metazoan genomes. In my opinion, the article is well written and easily understandable by non-specialists. I only have minor concerns that I would like the authors to address if they agree with me.

Introduction

7;894 = 7,894
The value was corrected.

Sequencing

Figure 1: I understand the intent of this figure, but I find it pretty challenging to read, and points hide other points. One way of fixing this would be to aggregate the data of each category per year and turn it into a boxplot. Figure 6: Same as Figure 1
The disadvantage of a boxplot would be that it would not highlight the transition between sequencing technologies. Besides, using points rather than boxplots puts an emphasis on their distribution (and outsiders).

The resulting reads have a length around twenty kilobases (kb): In my experience, PacBio reads usually have a mean size around 15kb that can go up to 25kb (see https://www.nature.com/articles/s41597-020-00743-4 as an example).
We modified this value.

The error rate has also been decreasing with the release of new flow cells and the development of more accurate basecallers such as Bonito. There is also a new protocol called Q20+, which makes it possible to generate reads with a 1% error rate.
We completed the text and added a reference on Q20 Nanopore reads.

Genome assembly

DBG-based assemblers require highly accurate reads in which errors are only substitutions, with no indels: why should there be no indels?
This statement was indeed not completely accurate and we modified the sentence.

To this end, heterozygous regions are collapsed in order to keep a single sequence for every region in the genome: this is true if the genome is not very heterozygous. In the other scenario, both haplotypes can often be retrieved, as heterozygous regions are pretty different.
Although it would make more sense to separate haplotypes for highly heterozygous genomes, it seems that the norm is still to collapse into haploid assemblies. We expect and hope for it to change though with highly accurate long reads, which is the purpose of the "Phasing assemblies" section.

Assembly pre and post-processing

Table 2 - Long reads error correction: NaS is missing. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1519-z
Thank you for this suggestion.

Table 2 - Short and long reads polishing: a new tool called HAPO-G has been published recently and is absent from the list. It has been developed explicitly to polish heterozygous genomes but also handles homozygous genomes. https://academic.oup.com/nargab/article/3/2/lqab034/6262629
We had mentioned Hapo-G in the phasing section, but we were not aware that Hapo-G was also suitable for haploid assemblies and we added it to the list.

Drawbacks of using Hi-C are not presented. As an example, the fact that gap sizes cannot be estimated is not indicated.
We now mention the problem of gap size estimation, as well as the limitation of the minimum input required by Hi-C protocols.

Assembly and pre/post-processing steps are often combined in one tool makes it look like there is no need to post-process assemblies further, but if the polishing step is only done with long reads, the final quality will not be great.
We modified the sentence to make it clearer.

Phasing assemblies

Hifiasm is another assembler that can phase haplotypes.
We added hifiasm to this section.

Reviewer 4

This work reviews the current state of methods for genome sequencing and de novo assembly, with a particular focus on invertebrates, for which resources are still missing. This sort of work should be encouraged, as it aims at expanding genomic resources to non-model species, which is crucial to obtain a more comprehensive picture of the evolutionary and mechanistic processes underlying biological diversity. The "technical" content is comprehensive and mostly up-to-date. My main concerns are mostly revolving around the structure and the scope of the review. In its current state, it reads like a rather "generic" review about assembly tools, with illustrations drawn from genomic studies of invertebrate species. I think that the review would benefit from a more explicit description of the specific challenges encountered in invertebrates. Low DNA amounts is mentioned, but there are other aspects that could be described. For example, many species are difficult to raise in controlled conditions, or rare in the wild, or poorly described from a taxonomic perspective. On the other hand, many species of arthropods reproduce asexually (e.g. Daphnia), which may help increasing the yield of DNA from the same genotype. At the moment, it reads more like a collection of anecdotes (which I agree all reveal an interesting problem): there may be a better way to structure it.
It would also be good to explain from the beginning the readership that this review targets. For example, I understand the interest of adopting a historical perspective in the first section (Sequencing) if the review is a resource for new practitioners. However, a review that aims at explaining the current methods for genome assembly to "naive" readers should take more time explaining basic concepts (e.g. N50). A glossary could be useful. On the other hand, if the review is addressed to scientists who already have some experience with the techniques and the terms, the somewhat long description of Sanger sequencing may not be particularly useful.
In my opinion the review does not provide (yet) a guide to decide of a sequencing strategy. The information is already there, but could be highlighted in a more organized way. Figure 6 is a good example of what could be done more extensively throughout the review in my opinion (with more details).
We developed the introduction to better describe challenges specific to non-vertebrate animal species and the target of this paper. We also added a new section "Recommendations" to provide guidelines for genome assembly.

The authors could compare the quality of currently available assemblies, using several metrics, and highlight the methods used to obtain them. For example, what sequencing depth of coverage is needed when using only Illumina reads + mate pairs? Hi-C? PacBio + Illumina short reads? What is the average cost? It would be useful to have figures such as decision-making flowcharts. Figure 5 could be expanded to highlight the different possible options at each step (short-reads? Long-reads? What is the best option given a budget of 10,000$? 50,000$?). What are the bioinformatic resources needed? What is the runtime of different programs, and how this runtime scales with genome size and complexity?
Providing guidelines on the sequencing depth required for assemblies with Illumina reads, HiFi reads, or scaffolding with Hi-C could indeed be useful, however, they could constitute three different papers on their own. In a previous paper (https://doi.org/10.1186/s12859-021-04118-3), we tested 7 assemblers of low-accuracy long reads on a small eukaryote genome with different sequencing depths, and we measured RAM usage and CPU time. But this analysis, restricted to low-accuracy long reads on one genome, represented 1.5 years of work, and conducting a thorough analysis of all sequencing methods would certainly represent a massive amount of work which would be outdated by the time it is completed.

I also think that mentioning reference-guided assemblies could be useful, especially for readers who consider working on a species related to one that has already been sequenced. If there are reasons to assume that synteny is high and divergence low, reference-guided assemblies may be a good way for researchers with limited financial resources to obtain a valuable resource. A particularly interesting paper from this perspective (in my opinion) is the following one (Lischer Shimizu, BMC Bioinformatics, 2017):

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1911-6

Note that this paper also proposes an interesting way to test for the quality of assemblies obtained by different methods through the combination of 36 summary statistics (using z-scores for each of the statistics and comparing their distributions across methods).

P17: Assembly evaluation. There are so many ways to estimate the quality of an assembly that some authors have proposed a set of tens of summary statistics, that they summarize as a Z-score. (check papers on reference-guided assemblies).
The paper is oriented towards non-vertebrate animal species, and more specifically non-model ones, for which references are not yet available. That is why we focus on reference-free evaluation methods, as explained in the "Assembly evaluation" section: "For genome assemblies of non-model non-vertebrate animals, reference assemblies are seldom available, or they have a poor quality or contiguity that the new assembly aspires to improve. Therefore we will focus on reference-free evaluation methods.". There has been several metrics published that aimed to combine quality metrics. However, combining metrics tends to hide specific issues in an assembly, that may impact more or less the conclusions that can be drawn.

At last, it may be worth explaining what can be done with a genome assembly depending on its quality. If the goal consists in running preliminary population genetics analyses, a fragmented assembly can already be very useful. For comparative genomic analyses, assessment of repetitive content (transposable elements), or functional studies, high quality assemblies are the target to reach.
We now mention this problematic in the section "Recommandations".

Nevertheless, I want to emphasize the fact that the review is rather comprehensive, and mostly needs polishing to increase its impact on a broad range of readers.

Minor comments through the text:
Introduction, Paragraph 6: The bit about BUSCO feels slightly too long, although the issue highlighted is very interesting. There are many other possible biases that could be discussed. Maybe shorten it, and provide other examples of how bias towards model systems can impair research on non-vertebrates. In general, the Introduction would benefit from explicitly stating the scope of the review, and what it means to achieve (decision-making tool? Comparison of methods? Introduction to the field for new practitioners?).
We completed the introduction.

Sequencing, second paragraph. N50 is usually low for second generation sequencing, as you mention, but using Hi-C, Hi-Fi or mate pairs (which I would still classify as second-generation sequencing) can improve assemblies a lot.
Hi-C and mate-pairs indeed include a short-read sequencing step, but the long-range information they provide is used for post-processing with scaffolding. When they are used for *de novo* assembly, their specificities of Hi-C reads or mate pairs are disregarded. In the case of HiFi reads, the technology is quite different from second-generation sequencing, although similar in accuracy, but it may fit in its own category of fourth-generation sequencing, with high-accuracy long reads.

Sequencing, third paragraph. The current increase in accuracy for base calling and assembly from nanopore reads is encouraging, but should be discussed more in terms of minimum depth of coverage required, the quality of training datasets (for algorithms using machine/deep learning), etc. Note the existence of another base-caller, Poreover, to be used in combination with Bonito https://github.com/jordisr/poreover
We expect that training datasets are more adequate for species for which resources are already available, but our species is rather on non-model species. We added Poreover to the text.

Table 1: This table is a good resource, but it may be worth considering merging it with table 2. A classification highlighting speed and memory requirements would be useful. As mentioned in the main comments, I am not sure that the row on first-generation sequencing is particularly useful.
We believe that combining Tables 1 and 2 would make them harder to understand. A thorough benchmark could indeed be useful. However, values provided in papers are not necessarily comparable as they may be obtained on different machines with different datasets, and with less optimized versions than the ones currently available.

P8: You talk here about k-mers, but what about decisions on which k-mer length to use? Why is it important to use several k-mer lengths when assembling? This is something that you could already explain here.
We completed the text.

Figure 4: It would be interesting from a decision-making perspective to add a panel with the different techniques used to assemble these genomes.
The techniques used for assembly are not indicated in Figure 4, but they are presented in Figure 1.