Diving, and even diging, into the wild jungle of annotation pathways for non-vertebrate animals

Francois Sabot

Peer Community In Genomics

Diving, and even diging, into the wild jungle of annotation pathways for non-vertebrate animals

Francois Sabot *based on reviews by Yann Bourgeois, Benjamin Istace, Cécile Monat and Valentina Peona*

A recommendation of:

A deep dive into genome assemblies of non-vertebrate animals

Nadège Guiglielmoni, Ramón Rivera-Vicéns, Romain Koszul, Jean-François Flot *(2022), Preprints, 2021110170, ver. 3 peer-reviewed and recommended by Peer Community in Genomics* ***https://doi.org/10.20944/preprints202111.0170.v3***

Open Access

Submitted: 10 November 2021, Recommended: 21 April 2022

Published: 06 May 2022

Cite this recommendation as:
Francois Sabot (2022) Diving, and even diging, into the wild jungle of annotation pathways for non-vertebrate animals. *Peer Community in Genomics, 100016.* ***https://doi.org/10.24072/pci.genomics.100016***

Recommendation

In their paper, Guiglielmoni et al. propose we pick up our snorkels and palms and take "A deep dive into genome assemblies of non-vertebrate animals" (1). Indeed, while numerous assembly-related tools were developed and tested for human genomes (or at least vertebrates such as mice), very few were tested on non-vertebrate animals so far. Moreover, most of the benchmarks are aimed at raw assembly tools, and very few offer a guide from raw reads to an almost finished assembly, including quality control and phasing.

This huge and exhaustive review starts with an overview of the current sequencing technologies, followed by the theory of the different approaches for assembly and their implementation. For each approach, the authors present some of the most representative tools, as well as the limits of the approach.

The authors additionally present all the steps required to obtain an almost complete assembly at a chromosome-scale, with all the different technologies currently available for scaffolding, QC, and phasing, and the way these tools can be applied to non-vertebrates animals. Finally, they propose some useful advice on the choice of the different approaches (but not always tools, see below), and advocate for a robust genome database with all information on the way the assembly was obtained.

This review is a very complete one for now and is a very good starting point for any student or scientist interested to start working on genome assembly, from either model or non-model organisms. However, the authors do not provide a list

of tools or a benchmark of them as a recommendation. Why? Because such a proposal may be obsolete in less than a year.... Indeed, with the explosion of the 3rd generation of sequencing technology, assembly tools (from different steps) are constantly evolving, and their relative performance increases on a monthly basis. In addition, some tools are really efficient at the time of a review or of an article, but are not further developed later on, and thus will not evolve with the technology. We have all seen it with wonderful tools such as Chiron (2) or TopHat (3), which were very promising ones, but cannot be developed further due to the stop of the project, the end of the contract of the post-doc in charge of the development, or the decision of the developer to switch to another paradigm. Such advice would, therefore, need to be constantly updated.

Thus, the manuscript from Guiglielmoni et al will be an almost intemporal one (up to the next sequencing revolution at last), and as they advocated for a more informed genome database, I think we should consider a rolling benchmarking system (tools, genome and sequence dataset) allowing to keep the performance of the tools up-to-date, and to propose the best set of assembly tools for a given type of genome.

References

1. Guiglielmoni N, Rivera-Vicéns R, Koszul R, Flot J-F (2022) A Deep Dive into Genome Assemblies of Non-vertebrate Animals. Preprints, 2021110170, ver. 3 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.20944/preprints202111.0170

2. Teng H, Cao MD, Hall MB, Duarte T, Wang S, Coin LJM (2018) Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience, 7, giy037. https://doi.org/10.1093/gigascience/giy037

3. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105–1111. https://doi.org/10.1093/bioinformatics/btp120

Reviews

Toggle reviews

Reviewed by Cécile Monat, 23 Mar 2022

This new version of the manuscript is enriched with corrections and precisions to anwered me and other reviewers questions and suggestions, making it a better article.

Reviewed by Yann Bourgeois, 11 Apr 2022

The authors have provided an answer to all my main comments, and I mostly agree with them. I still believe that a rough estimate of the current (and past) costs for different techniques could be provided, acknowledging that this may become quickly obsolete. This would provide an upper range estimate for teams intending to start a genome sequencing project. The review is already thorough and I am happy to support acceptance.
Best wishes,

Yann Bourgeois

Reviewed by Benjamin Istace, 15 Mar 2022

The authors have successfully addressed all my concerns.

Evaluation round #1

DOI or URL of the preprint: *10.20944/preprints202111.0170.v1*

Version of the preprint: 1

Author's Reply, None

Download author's reply

Decision by Francois Sabot, 06 Jan 2022

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.

Sincerely yours,

Dr Francois Sabot

Reviewed by Cécile Monat, 29 Nov 2021

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

Reviewed by Valentina Peona, 17 Dec 2021

Download the review

Reviewed by Benjamin Istace, 26 Nov 2021

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

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.
«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).
«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.

Genome assembly
«DBG-based assemblers require highly accurate reads in which errors are only substitutions, with no indels»: why should there be no indels?
«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.

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
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
Figure 6: Same as Figure 1
Drawbacks of using Hi-C are not presented. As an example, the fact that gap sizes cannot be estimated is not indicated.
«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.

Phasing assemblies
Hifiasm is another assembler that can phase haplotypes.

Download the review

Reviewed by Yann Bourgeois, 17 Dec 2021

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).

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?

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).

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.

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?).

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.

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

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.

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.

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).

Figure 4: It would be interesting from a decision-making perspective to add a panel with the different techniques used to assemble these genomes.