2023.08.29.555262v2 review

The study really nicely introduces the impact and significance of producing a high quality, chromosomally scaffolded assembly of the brown hare, describing in detail its population changes through expansions and contractions, as well as colonizing non-native environments, and hybridizing with other hare species. The publication of this genome reference will serve as a perfect base for future population genetic studies in hare populations and species boundaries studies. The study presents a genome assembly produced with long read data and Hi-C, that covers the 23 autosomal chromosomes, X and Y. The study also presents the mitochondrial genome of the brown hare.

The study needs some work on the material and methods, given the main part of this study is the sequencing and genome assembly process, as well as some better introduction to the methodology. It is also missing a bit of cohesiveness between results and discussion.
I believe this article would be ready for publication upon careful revision of the following points mentioned below:

List of comments and suggestions
Lines mentioned based on the PDF in biorxiv: 2023.08.29.555262v2.full.pdf

Introduction
Line 52: bit it remains as native species (suggestion, but anything that helps the sentence flow better would be good)

Line 58: especially through the expansion...

Line 65: Are the mountain hare and the SA Cape hare the same species? If not, specify the species name of the mountain hare, as done for the SA Cape hare.
Extra note: Mountain hare is then mentioned in line 86 again, and the species name specified after, it should instead be changed to the first occurrence (line 65).

Line 73: Maybe link better the previous sentence with: Identifying Poland as.. at first, it is unclear why only that scenario for type locality is explained.

Line 111: There is an ongoing discussion regarding the continuation of utilization of the terms second or third generation sequencing, as technologies are developing at such fast pace and very different technologies in parallel. I would suggest initiating that sentence by just saying, the technological development of high-throughput sequencing of long molecules, …
Same on line 118 (change to long read sequencing)

Line 114: early 2000's is very generic, and not clear what you mean by that, I would say that through a big part of most early 2010's it was still also mostly genome assemblies based on short reads. (For example, PacBio RS did not start being fully commercial until 2011. Oxford Nanopore's MinION until 2015)

Line 114: Same comment as before, about specifying second generation, I would just say short read sequencing technologies. Short read data is still in constant development and is still needed and useful for a lot of studies (including Hi-C sequencing, mentioned after).

Line 115: Instead of similarly, I would say additionally, as it is yet another technology that can be used supporting the others.

Line 115: Propelled these genomes -> it would be more accurate to say that has propelled the research on generation of genome assemblies and reference genomes

Line 117: Weird phrasing: I would rephrase to "Coupled with these advances, the decrease in price per base pair through the years has made whole genome sequencing available to many laboratories and research groups"

Line 120: these two technologies: specify there which long read technology, as it is the first time mentioning it in the main text.

Line 133: Even if it is quite clear by context, this paragraph is important and I would clarify again "a male specimen of brown hare". The previous paragraph talks about different species and genome assemblies, so I think it is good to specify again that the study is on brown hare.

Line 137-146: It is good to give final results on the final paragraph of the introduction, but I think it is too detailed, and most things should be left for the actual results and discussion sections. For example, no need to specify each BUSCO score here.

Line 142: I believe you mean L90 and not N90 there.

Material and Methods

Overall, revise proper referencing of machines, protocols, and materials used

Line 189: Is there no reference to that protocol?

Line 190: reference properly the machines used, Qubit should also have information about the company and country in parenthesis (ThermoFisher, Country).

Line 191: The DNA was not analyzed, it was sequenced.

Line185-193: Who did the library prep? Which library prep was used? Add information on this, as it is key information in this paper.

Line 195, section Mitochondrial DNA. Just out of curiosity, was the mitochondrial genome not found among the PacBio raw data? Even if not assembled, it is usually possible to find circularized reads that perfectly cover it. It could be a good way to test that the separate method gives a similar result.

Line 264: Specify company or reference information on NEBNext Ultra…

Line 267: Ampure beads, again, company missing

Line 274: it is not clear to me why those arguments are necessary to integrate the Hi-C data, if that is done after, separately from that first genome assembly. Could you describe more clearly that part of the methods?

Line 280: What was used to sort and deduplicate?

Line 298: there is no explanation on how the RNA sequences were assembled. In line 134 you mention that RNA was also extracted, but then there are no methods on how it was sequenced, and then assembled, as mentioned later in the genome annotation section. If it has been published in a previous publication, there is also no reference to that.

Line 316: Does Geneious have a reference? If not, the official website can be an alternative to that.

Line 320: I am aware that manual curation is very difficult to describe, and it ends up becoming like this method black box in the end, but it would be nice to know what level of manual curation was done. It is not the same to fix a couple of wrongly annotated genes, compared to rearrange contigs and scaffolds. It would be nice to just mention an approximate description of how much manual work it was done in this precise genome assembly.
I saw after that in lines 352-353 it is described more in detail; I would appreciate something like that in the methods. Just specifying the type of curations that were made.

Results

Line 332-334: I would add this at the end or beginning of the material and methods section. But this brings up another problem: in the methods, only the primary assembly has been described, and nothing about an alternative assembly. The methods should include all information about how the results have been produced!

Line 338: estimated, not published

Line 341: using a k-mer size of 21 (no parenthesis needed). Maybe the sentence as a whole needs rephrasing.

Line 347: the largest scaffold is the same as before, so I would phrase it better to refer to that back. Right now, it seems like it is different.

Lines 337-351: Be a bit clearer on the differences between the two assemblies. Right now, it is just a lot of sentences giving numbers, but I am missing some clear distinction and similarities between them. I do not mean as a discussion, but more to make it clear to the reader. When the Hi-C scaffold assembly starts being described, again all the numbers are mentioned and it is hard to remember what was the same or what has changed from the previous. I would revise the entire paragraph to ease the reading, while still maintaining all the important numbers and data.

Line 352: Hi-C maps is not mentioned in material and methods, and it is not referenced in this first time that comes up.

Line 356: 93.16% what? Contigs? Scaffolds? Total Mb?

Figure 2: A -> Genomescope2 is not mentioned in material and methods nor properly referenced. B -> What was used to create the Hi-C map? C -> even if explained in methods, and noted on the axis, add explanation of what is the previous assembly.

Line 370: You keep referring to the previous genome assembly, but not clarifying that it is not the same hare species.

Line 368-381: Has anyone checked if there is multicopy genes in the repetitive element libraries? It is something worth doing, as you might be hiding important host genes in the masked section of the genome assembly.

Line 385: refrain from using “genome”, use genome assembly instead.

Line 389: there are*

Line 395: in the assembly*

Line 395: Why only discuss a fusion in the rabbit and not a chromosomal split in the hare?

Line 397: again, genome, and then reference genomes. Specify as genome assemblies.

Line 402 answers my previous question on the mitochondrial data. But I am surprised that no comparison with the actual sequence from HiFi is not done. It could be a great way to test if your sequence is correct.

Discussion

Too many times throughout the discussion to note them down: change genome to genome ASSEMBLY.

Line 439: If my comment above is to be investigated in future studies (possibility of host genes in the repeat libraries), mention in the discussion that possibility, so that it is clear that it could be done in the future. Right now, I would either suggest a major revision on that, or being very clear about being aware of the possible problems that might have caused that higher level of repeats in your genome assembly.

Line 463-467: consider rephrasing those sentences, it is a bit confusing as it is now, with very long sentences and not good flow.

Line 467: Heteroplasmy already commonly denotes more than one mitochondrial genome variants, with no specification of two or more. Instead of saying “multiple heteroplasmy”, it

would be better to add the clarification of what you mean by that after mentioning heteroplasmy.

Line 470-482: It would be nice to add some references here. Including methods and the mentioned law.

The discussion is a bit weirdly structured. It starts with a very short paragraph of the actual results here presented, followed by a long explanation on the mtDNA, and a big section on DNA obtention for these kinds of results. Meanwhile, during the results, there were many parts that could be considered discussion, with references added to what was being presented. I would suggest moving some of that information to the discussion, to properly discuss the results that are presented in the actual study.