Review by anonymous reviewer 1, 20 Jun 2024 08:20

> The authors have addressed my comments.
>
> In the response, they provided the sentence "Long indels are treated as multiple adjacent loci." They could also add it to the main text to make it clearer.

This has been added.

Review by Dmitry Antipov, 08 Jul 2024 20:09

> During the revision authors made great job on improving both text and tool itself.
> However there are still some moments where improvement is possible:
>
> Major:
> Results still barely mention anything except completeness (I found only one sentence "Particularly with Nanopore data, HairSplitter produced the most complete assemblies, though less contiguous than those produced by Strainbery."). Focus on the completeness metrics is clear, but both correctness and contiguity deserves more attention. I.e. there are hundreds of misassemblies for some datasets in supplementary table 3, and if they are because of the regions of different strains assembled into chimeric contigs this definitely can affect downstream analysis and should be mentioned in the Results or Discussion
>
> **We added this paragraph in the Results section :**
>
> **The accuracy of the contigs produced by HairSplitter was found to have a lower number of indels and mismatches compared to iGDA and Strainberry (Sup. Tables 2 and 3). This confirms that the groups of reads used by HairSplitter to build the contigs were more homogenous in terms of haplotypes. However, all tools produced a significant number of misassemblies when reconstructing a high number of strains. In the case of HairSplitter, these misassemblies were primarily caused by the fact that a few small structural variations were not detected in the graph completion step. In terms of contiguity, all assemblers produced comparable results, although HairSplitter appeared to make slightly more conservative choices than Strainberry, resulting in a slight decrease in contiguity but a lower number of misassemblies (Sup. Table 2 and 3).**
>
> Minor:
> Possibly this is biorxiv bug but still - there's some mess whether tables 2-5 are suplementary or not. pdf version is consistent, and text refers them as suplementary but web https://www.biorxiv.org/content/10.1101/2024.02.13.580067v2.full shows them in the main text.

This has been corrected.

Line 118: contigs of the _completed_ assembly? Likely so, but not 100% clear

Yes, we inserted the word completed

Line 336: high duplication ratio reference to sup table2 can be beneficial here

This has been added

Line 294-295: phasing of polyploid organisms
Cited paper was published before the current age of T2T assemblers (hifiasm, verkko) and do not distinguish diploid and polyploid organisms. That assemblers do not have significant problems in separating haplotypes (for polyploids there's still a problem with phasing but in different sence - utilization of Hi-C or other long distance technologies and not in the long read level). So additional motivation for extending hairsplitter to polyploids would be beneficial.

We modified the paragraph and updated the citation, citing as a motivation a list of completed polyploid plant genome as of 2023: none of them was assembled using exclusively long noisy reads:
Since HairSplitter is already successful at separating both bacterial and viral haplotypes, we expect to be able to extend this work naturally towards the phasing of polyploid organisms, motivated by the fact that for now polyploid genome assembly requires highly precise illumina or HiFi reads (Kong et al. 2023)

Lines 255, 258: suplementary table 4 instead of 5? Also suggest to add mention about Strainline crash to the caption of that table too.

This has been added.

metaMDBG removal - I get the motivation, but since it is a popular tool would be nice to explain it in the text too

We now explicit “Software that purposefully collapse similar strains, such as metaMDBG \ cite{metamdbg}, have been left out of the benchmark.”