Author's Response To Reviewer Comments

Close

Journal: GigaScience
Manuscript ID: GIGA-D-20-00061
Title: " Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford Nanopore"
Author(s): DanDan Lang; Shilai Zhang; Pingping Ren; Fan Liang; Zongyi Sun; Guanliang Meng; Yuntao Tan; Jiang Hu; Xiaokang Li; Qihua Lai; Lingling Han; Depeng Wang; Fengyi Hu; Wen Wang; Shanlin Liu

Dear Dr. Hans Zauner,
We are very grateful to both the reviewers and the editor for the critical comments and constructive suggestions, which have helped improve our paper considerably.
Below we provide our responses to the reviewers' comments in blue. We have incorporated most of the suggested changes as well as additional analyses. The manuscript has now gone through required revision and reorganization, and we sincerely hope that this revision is satisfactory to the reviewers

Reviewer #1: This manuscript compares the results of genome assemblies from the data of two long-reads sequencing technologies and multiple genome assemblers. It focuses on analyzing the impact of the sequence qualities (read lengths and accuracies) to the contiguity and the accuracy of the assembled contigs.

While the results agree with the general understanding of how the read lengths and the basecall accuracies affect the final assembly quality, I found the detailed examples comparing the two picked assemblies are interesting. It provides useful insight for understanding the impact of repeats for genome assembly results for researchers. The manuscript is well written and easy to follow to get the points across. Here are a couple points that I hope the authors will be able to address:

(1) While the rice strain is documented in the manuscript, it will be useful to comment on the polyploidy of this particular strain? The BUSCO results seem to indicate it is a haploid strain, and the readers may be able to check it out from the strain ID. However, the authors should comment on the polyploid to help the readers. It is important to understand how to interpret results according to the known polyploidy.

> The rice individual (Oryza sativa) we used in this study is the indica cultivar 9311, which is a diploid strain. We noted it at line 70.

(2) In the paragraph starting with "Following DNA extraction", please refer to the supplementary material about the extraction protocol there.

> To conform to the journal style, we moved part of the supplementary methods to the main text, which should have solved this problem. Thank you for pointing it out.

(3) The authors should comment on the time used for sequencing on PromethION and Sequel II, and the computation resources (CPU/wall clock time, memory, cluster setup, etc.) needed for each assembler.

> It is a good suggestion. We included it at Table S1 in the resubmitted version.

(4) The IGV view of the ONT reads mapped the PacBio assembly GAP doest show the disagreement of the ONT reads to the ONT contigs. While the high error rates may it messy to see. If such a view hard to see, it still useful to examine if there is some systematic disagreement between the reads and the contigs. I am hoping the authors can comment on whether some systematic errors are visible. Also, will it provide useful insight if we compare it to PacBio Reads mapping to the ONT contigs?

> The IGV plot aims to demonstrate the GAPs of the PB HiFi assembly can be spanned by several ONT

ultra-long reads, and thus explained the reason why such gaps can be assembled using ultra-long reads. Zoom in the IGV plot may show the systematic errors. However, it will as well dismiss our main purpose. Therefore, we would keep it as its current view.

(5) When the authors refer to "string graph," it needs a citation. The term the "string graph" is coined by Gene Meyer for a specific way to construct a graph for genome assembly. Not all assemblers use the same graph construction. The authors should use "assembly graphs" and cite related papers.

>We added the corresponding citation, and algorithms of the software referred to here is based on string graph, so we kept the term “string graph”.

(6) Related to the polyploidy of the strain, the author mentioned "diploid heterozygous states," there is no citation or explanation to help the readers to know what the authors refer to.

>As assembly obtains one single suite of a diploid genome, only one state of those heterozygous sites presents in the assembly results. The differences between the ONT and the PB assembly could be the real conditions in the individual we sequenced. We clarify it at lines 230-231.

(7) The authors mention the errors in ONT assembly are clustered. The authors' explanation is because of low coverage mapping in the polish steps. Are these clusters caused by repeat contents, low accuracy of ONT assembly op particular sequencing contexts? In the caption of Figure S5, the authors write: "the distances should have a peak around 1,000 bp for an average error rate of 1.06 per kb in the case of random distribution." The author should put a theoretical curve or a simulated one on the same plot to show the distribution of a random error model does generate a different distribution.

>Thank you for the suggestion. Reviewer #2 also proposed a similar suggestion. We further investigated the genomic characteristics in and flanking those error regions. It showed that those error-enriched regions were characterized with higher methylation level compared to the other genome regions, and we added it at lines 146-150. We also added a theoretical curve on Figure S6 (Figure S5 in the last version) to better illustrate our point of view. Thanks for this constructive suggestion.

Reviewer #2: In the manuscript entitled, "Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of PacBio Sequel II and ultralong reads of Oxford Nanopore," Lang et al., generate assemblies for a rice variety (9311) using the two different long read sequencing technologies and then compare contiguity and accuracy statistics. The authors conclude that Oxford Nanopore Technologies (ONT) sequencing provides superior contiguity, while Pacific Bioscience (PacBio) provides superior base quality accuracy, and that the two platforms should be leveraged together for reference quality genomes. Overall the manuscript is very concise and well developed. However, there are a couple points that the authors should acknowledge and discuss, which impact the interpretation of their results.

First off, the BioProject PRJNA600693 was not available to assess the assemblies or the raw data. In a manuscript that compares genomes, validating some of the claims is essential, and the data should be available to the reviewer.

>Thank for pointing it out. It is assessable now. In addition, to follow the rule of GigaScience, we have already uploaded the two assembly files, two annotation files, two complete BUSCO output files, two CDS sequence files, two protein translation files and alignment results to the GigaDB server in the process of our first submission. It should be available to reviewers. The access info is as follows:
username = user30
password = LiuSComparison
FTP server = parrot.genomics.cn

The authors set up a very nice and simple contrast between PacBio HiFi and ONT. There are some significant differences between the datasets that should be discussed though. The read N50 length of the two platforms is considerably different at 41 kb vs. 13 kb for ONT and PacBio respectively. Moreover the absolute coverage is significantly different between ONT and PacBio at 92 Gb (230x) vs. 253 Gb (632x) respectively, even though the reported HiFi coverage is only 50X. There are several opportunities here. First, the authors should at the very least mention these differences, which at face value explain ONT being more contiguous and PacBio having higher base quality. Second, since the authors have an extraordinary amount of data for this rice line, it would be also interesting to see where the quality or

contiguity starts to decrease as a function of the amount or type of data.

>Good point. We clarified the coverage differences between the two platforms in the resubmitted version at lines 77-81. We also subsampled the raw reads to investigate the influence of data size on genome assembly, please find it at lines 127-130 and in Table S6 and Figure S4.

The section about the nucleotide variation is a little confusing. It is stated that the regions (~94%) that showed low base quality in the ONT assembly also had low shotgun read coverage. Was this ONT, PacBio or Illumina coverage that was low? With the amount of coverage that was generated for each platform (ONT, 230x; PacBio, 632x; Illumina 70x) why would there be regions in the assembly with less than 5x coverage. This needs to be clearer. In the same section, SNPs and INDELs are referred to as small-scale mis-assemblies; more accurately these are sequence errors not mis-assemblies. Did the authors use the ONT or PacBio data to look at DNA methylation? If the errors are clustering in the genome then maybe the errors in the ONT sequence are the result of mis-called bases that are highly methylated. Since the data is available this would be an important point to make or reason to rule out.

>It is a very good point regarding to the abnormal coverage issues. Firstly, we clarify that the low coverage refers to the shotgun reads generated using MGI-SEQ platform. Then, we added possible reasons that deterred the correct mapping of short reads for those regions, please find them at lines 146-156.
For the word “mis-assembly”, we agree that those SNVs and InDels should come from sequence errors. We clarified it at line 140.
It is a good suggestion as for the DNA methylation analysis. We investigated the correlation between methylation profiles and those error-enriched regions. It showed that the GC content and methylation level of those error-enriched regions are significantly higher than that of other genome regions. We included it at lines 150-156 and Figure S8.

PacBio can also run in long read mode, so researchers could mix HiFi with longreads on one platform. This would be good to also mention.

>Added, at lines 128-132.

The BUSCO scores for the two genomes are almost identical. It would be good to add a bit of commentary why you see similar BUSCO scores but some differences in protein content. This will help the reader understand the differences and limitations of each measure.

>Thank you. We included the explanation at lines 153-156.

While mentioning exact costs for both methods would not stand the test of time it would be good for the reader to understand the relative cost differences between the two approaches.

>Since the yield of both the platforms (especially the ONT) varies a lot between different species. For example, some human DNA samples can generate > 100 Gb data using one PromethION cell, but some marine or insect species can only generate < 20 Gb data per cell. As a result, we don’t think cost of the current work (both platforms have spent around $4,000 for sequencing) reflects a real cost difference for other species. It would be better for the readers to consult their local dealers for the cost details.

Minor points:
What species of rice is 9311? The authors should use the scientific name somewhere in the manuscript to clarify what species is "rice."

>We corrected it as “one rice individual (Oryza sativa indica, 2n = 2x = 24, variety 9311)” at line 69.

Grammar:
The first sentence of the Main Text. Diseases don't find causative alleles. Maybe, "The human reference genome enabled the identification of disease causative alleles…."

Sentence 4 page 3: species don't leverage cutting edge sequencing. "The two cutting edge sequencing technologies has enabled the sequencing of many species…"

Bottom page 4 "It was gone through by multiple ONT long reads…" It was spanned by….

>Thank you for noticing those errors. We have them corrected accordingly.

Reviewer #3: Advances in sequencing technologies provide us with an unprecedented opportunity for high-quality de novo reconstruction of complex eukaryotic genomes.
The manuscript presents the comparative analysis of the two assemblies of a rice genome, obtained with ultra-long ONT and Pacbio HiFi sequencing.

First, while a combination of HiFi and ultra-long ONT datasets is available for several human genomes (and maybe some other organisms), the scope of the study is limited to a single organism with a relatively small and simple genome. Moreover, only a single genome has been considered with a single dataset for each technology. In particular, while longer Pacbio HiFi libraries with reads reaching 20Kb are now not uncommon the dataset considered in the study had an average read length less than 12Kb.

>Firstly, human genome, as well as model species, could be special cases. For instance, scientists who work in the field of human health could account for more than half of the entire academic world. They depend heavily on one single genome reference and have been spending tremendous time and money to achieve high-quality genome references, and thus combined as many cutting-edge technologies as possible. However, the vast majority of scientists who study non-model species obtained the genome references of targeted species using only one single sequencing tech, either PB or ONT, due to limited funding. The current work provides scientists valuable information on the pros and cons of PB HiFi and ONT ultra-long, and thus help them decide which one fits their project better, and they can as well learn the disadvantages of their choices in advance, as a results, be cautious to any related conclusions.
>For the library size, more and more studies begin to build long CCS libraries (15 kb – 20 kb) nowadays. We started this work right after the launching of PB sequel II. 10 kb library was recommended to guarantee high accuracy level for each CCS read at that time. We have an average HiFi read length of 13.36 kb, instead of what you mentioned: less than 12 kb which is the average length of subreads. We removed this confusing statement in the main text. In addition, we also added a note in the manuscript clarify this problem saying that "It is also worth noting that PB can run in long read mode, which, although can hardly generate reads as long as the ONT ultralong reads, can aid in connecting some of the gaps caused by long repeats. Besides, longer PB libraries with HiFi reads reaching 20 kb would be conducive to assembly contiguity as well".

Further I will focus on major issues of the presented analysis and mention some of the minor ones in the end.

Major issues

The 'primary' ONT assembly used was produced by a software tool for which I was not able to find neither publication/white-paper, nor a comprehensive benchmark. Moreover its github page states "In addition, we found that NextDenovo, of the current version, might produce a small number of unexpected connection errors in the highly repetitive regions, which, however, can be easily corrected using additional Hi-C or Bionano data. We are still in a progress of optimizing NextDenovo and will continuously update it, especially in terms of assembly accuracy". Since the only criteria used to choose the 'optimal' assembly between different assembly tools was based on their N50 values, it immediately raises questions about the reliability of the results!
The only confirmation of assembly accuracy given is the dotplot against the reference genome. Unfortunately at the presented resolution (of both the figure and the analysis itself) it fails to convince the reader of the structural accuracy of the assembly.
Also the discrepancy between N50 values of different ONT assemblies looks staggering and raises suspicion. I would suggest to include stats for some other well established long-read assemblers (e.g. Flye and Shasta), which will hopefully be able to produce assembly with continuity comparable to NextDenovo and dispel the suspicion.
As a side note, somehow the main text never states which assemblies were used for the most part of the analysis.

>NextDenovo is publicly available and free for downloading on Github. Up to the time we drafted this response letter, it has more than 2,000 downloads and eight releases (we used version 2.0 for this manuscript and the latest release is version 2.2). It is weird that the reviewer argued about the reliability of its assembly results because it generated a much better results compared to the other software. It is worth noting that its readme text on github states that it performs well especially for ONT ultra-long reads. It means the software developed algorithms to take advantage of ultra-long reads, just

like HiCanu designed its algorithms to fit HiFi reads. In addition, HiCanu also showed ca. 10 times higher N50 compared to the other two software. The discrepancy between HiCanu and the other two software for HiFi reads is almost the same to that of NextDenovo for the ONT ultra-long reads (10.38 vs 10.29). As both HiCanu and NextDenovo are publicly available on Github and both have not been certified by peer review, we believe this comment reflect the reviewer's personal preference.

>Although we think that this comment has more to do with the reviewer's preference than the actual merit of the manuscript, we added multiple genome assembly results using three more software, FLYE, SHASTA and NECAT, to avoid the staggering N50 differences. In addition to the collinearity analysis for large-scale assembly errors, and SNP and InDels analysis for small-scale assembly errors, we further examined the median size discrepancies between ONT and PB assembly to credit the accuracy of this ONT assembly. We included the results at lines 160-164.

One of the most surprising points of the analysis is that the authors insist on interpreting 'redundancies' as 'misassemblies', which is not a common practice in the assembly benchmarking. While it is important to highlight that while dealing with diploid genomes one can expect to get higher redundancy from HiFi-based assemblies, which should hardly be considered an error as long as they truly represent one of the haplotypes. Besides heterozygous differences, another potential source of redundancies can come from the fact that most long-read assemblers produce overlapping contigs, so the higher the number of fragments the higher will be 'redundancy' from those overlaps. Overall, I don't think that any types of redundancies should be considered as a serious problem at the assembly side. If needed, both types of common redundancies described above can be more-or-less straightforwardly removed post assembly (e.g. purge_dups software), but most importantly they stem primarily from particular algorithm implementation rather than show a deficiency of a data type. For example I would expect Flye's assembly of HiFi data to get much lower redundancy values due to more aggressive settings toward masking heterozygous differences and output of 'bluntified' contigs. Last but not least, from the methodological point of view, while I'm still uncertain how 'redundant' regions were annotated, they have been certainly detected against the draft ONT assembly, which could contain 'collapsed' tandem repeats and other issues, potentially inflating the stats.

>Objection. We defined those redundancies as mis-assemblies as we intended to assemble one suite of the diploid genome. Practically, the assemblies can be chimeric of the two haploids, rather than containing both haploids in one single assembly file. Most of the current analysis tools are designed to make use of such a genome reference, especially in the field of comparative genomics, which is as well the reason why some software (e.g. purge_dups as you mentioned) are developed to remove those redundancies. For instance, those redundancies could lead to incorrect deductions and conclusions in the analysis for gene expansion and contraction.

>It is worth noting that, instead of generating a perfect genome assembly, we aimed to report our observations objectively based on typical genome assembly pipelines for each sequencing platform, from which the readers can easily find out the advantages and disadvantages of both sequencing platforms and then decide what following analyses should be performed to improve their work. The software developers can also learn directly from the results to improve the corresponding assembly algorithms to avoid those unwanted mis-assemblies.

>The reviewer suggested ONT assemblies could contain 'collapsed' repeats and other issues, so could inflated our estimation. First of all, this argument is intuitive and groundless. Secondly, we defined those redundancies very careful, as what we mentioned in our manuscript, we checked the depths of those potential redundancies and classed them as redundancies only in the case that a total depth < 60X and depth of each < 40X. In addition, we also manually checked several corresponding regions on the ONT assembly to make sure they are spanned by single long read.

Significant part of the main text focuses on the analysis of a handful of particular cases of contig 'breaks' in HiFi assembly. First, the choice of 3 gaps taken for deeper analysis (corresponding to chr6) is not explained and, considering how few of them are described, it is unclear how well they represent the general situation. Second, at least some of the analysis is questionable. For one of the gaps the manuscript states that "... the overlapping and the gap regions represented two elements of 15 kb and 48 kb in length that, although have only one copy on Chr. 6, can find their duplications on Chr. 5 (Figure S3). Repetitive elements with such lengths go beyond the typical length generated by PB CCS, therefore the right path can hardly be disentangled from complicated string graphs." At the same time on Figure

S3 the sequence identity for instances of both repeats is reported below 98.5%! Repeat instances of such a high sequence divergence are extremely unlikely to affect HiCanu results, so there must be some other reason for fragmentation of this region.

I would recommend exploring the mapping of the HiFi reads onto the hypothesized genomic sequence, since it has been recently observed that HiFi reads can exhibit depletion of coverage in the GA-rich microsatellite regions of the genome. Besides being responsible for some of the observed gaps in this particular assembly, deeper investigation of this topic could have a serious impact on the choice of technology for certain assembly projects.

>Firstly, the scaffold for comparison was randomly selected and we added it in our manuscript to avoid confusion. Secondly, the three breaks showed in the manuscript are the entire set of breaks possessed in the selected assembly scaffolds for comparison, rather than that we chose the three. We would like to emphasize that we conducted the comparison analysis without any deliberate purpose to take side in any sequencing platform.

>For the sequence identity issues, we reported the average similarity score for the entire repeat regions (IDY of about 98.5%) between ONT assembly and PB assembly. The local similarity score can be up to 100% for regions > 10 kb. We believe those local high similarity regions are to blame for generating those gaps and redundancies. We included the local similarity scores on Figure S3 to avoid confusion.

As a final major note I would like to highlight that the data used in the study doesn't seem to be available yet (query of the PRJNA600693 id doesn't return any results on NCBI web site). TODO review was hampered by this.

>Thank you for your reminding. It is accessible now. Please find details in our response to Reviewer #2.

Minor issues.

If I understood correctly, the coverage of HiFi data exceeded 500x (253 Gb of data for a roughly 400Mb genome). Since it far exceeds the typical coverage of sequencing projects that most assemblers (e.g. HiCanu) are tuned to, I would suggest to subsample HiFi data or use HiCanu 2.0 (which would perform subsampling automatically) for processing a dataset of such coverage depth.

>We fed Canu self-corrected CCS HiFi reads which has a genome coverage of ca. 50X.

The authors note that "the errors of HiFi reads may be enriched in sequences with particular characteristics, rather than completely random … which may exacerbate the above error statistics for the ONT assembly", suggesting that the rate of the indels in polished ONT assemblies can be noticeably overestimated. I doubt that it is the case though. While the same properties of individual HiFi reads have also been recently observed by other investigators, to the best of my knowledge the consensus quality still tends to be very high. At the same time, the authors can make a much stronger claim by straightforwardly estimating the rate of 'false positive' errors detected within the regions of high coverage of unambiguously mapped Illumina reads.

>Firstly, we did NOT make any strong claims here, we said “may exacerbate the above error statistics for the ONT assembly” instead of what you mentioned “suggesting that the rate of the indels in polished ONT assemblies can be noticeably overestimated”. Secondly, we observed those disagreements between ONT assembly and HiFi assembly, and as what we stated in the manuscript, we also reckon that HiFi reads are of high quality, so we deemed those disagreements (SNPs and InDels) as errors of the ONT assembly. However, as Figure S10 showed, Illumina shotgun reads supported ONT assembly for some those differences and we carefully investigated the subreads of each CCS reads and found out that many subreads also supported the ONT assembly. Such information provided by subreads, however, lost during the CCS process. As it is impossible for us to manually check all such cases, we made a statement that “may exacerbate the above error statistics for the ONT assembly”.

The statement "PB assembly contained more gaps in each chromosome compared to that of ONT" can not be correct, since before that authors say that there were 3 cromosomes fully assembled from HiFi data.

>Corrected.

I would suggest against direct attempts at polishing HiFi assemblies with Racon, since it might result in corrupting the correctly assembled sequence within repetitive regions.

>Racon can correct lots of InDel errors for the HiFi assembly. As a result, we decide to kept it and added a note to remind readers of such an issue in Figure S11.

Conclusion.
Expectedly, while less than 60 genes were affected by identified assembly problems in HiFi assembly (most by redundancies, which as I mentioned before for the most part are easy to mitigate), even after polishing with Illumina reads > 1000 genes were affected by indels in the reported ONT assembly. Setting aside all the above mentioned issues, the results suggest the conclusion that ultra-long ONT could work well for scaffolding HiFi-based assemblies in order to produce almost-perfect genomic reconstruction of inbred rice varieties.

Overall, the presented manuscript falls short of providing the comprehensive comparison of the two technologies for sequence assembly (which a reader expects from its title), but works as a case study of how their combination should be able to provide an almost perfect medium-complexity genome of low-heterozygosity.

>As what we replied above, instead of achieving a conclusion of which platform is better and how to obtain a perfect genome assembly, we aimed to report the assembly differences between the two recently released sequencing techniques and provided a reference for those scientists who aim to generate genome references using one of these two sequencing techniques or both. Given the fact that other reviewers found our manuscript to be clear and easy to follow, this comment also seems to reflect the reviewer's personal preference. We did suggest that genome assembly work should leverage both platforms in the next-to-last sentence of our manuscript. However, the reviewer should not draw such a conclusion based on this single sentence, as all the above results talked about comparisons between the two assemblies.

Last but not least I found some parts of the manuscript quite poorly written. Additional rounds of revisions are highly recommended before resubmission.

>Thank you for your suggestion. We carefully checked the English writing thorough out the manuscript.

Close