GigaScience

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
--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00061
Full 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
Article Type:	Technical Note
Funding Information:	
Abstract:	The availability of reference genomes has revolutionized the study of biology. Multiple competing technologies have been developed to improve the quality and robustness of genome assemblies during the last decade. The two widely-used long-read sequencing providers – Pacbio (PB) and Oxford Nanopore Technologies (ONT) – have recently updated their platforms: PB enables high throughput HiFi reads with base-level resolution with >99% and ONT generated reads as long as 2 Mb. We applied the two up-to-date platforms to one single rice individual and then compared the two assemblies to investigate the advantages and limitations of each. The results showed that ONT ultralong reads delivered higher contiguity producing a total of 18 contigs of which ten were assembled into a single chromosome compared to that of 394 contigs and three chromosome-level contigs for the PB assembly. The ONT ultralong reads also prevented assembly errors caused by long repetitive regions for which we observed a total of 44 genes of false redundancies and ten genes of false losses in the PB assembly leading to over/under-estimation of the gene families in those long repetitive regions. We also noted that the PB HiFi reads generated assemblies with considerably fewer errors at the level of single nucleotide and small InDels than that of the ONT assembly which generated an average 1.06 errors per Kb and finally engendered 1,475 incorrect gene annotations via altered or truncated protein predictions.
Corresponding Author:	Shanlin Liu
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	DanDan Lang
First Author Secondary Information:	
Order of Authors:	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

| Order of Authors Secondary Information: | |
| Additional Information: | |

Question	Response
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our [Minimum Standards Reporting Checklist](). Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite [Research Resource Identifiers]() (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our [Minimum Standards Reporting Checklist]()?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or	Yes

deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the "Availability of Data and Materials" section of your manuscript.

Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?

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

Dandan Lang[1,#], Shilai Zhang[2,#], Pingping Ren[1], Fan Liang[1], Zongyi Sun[1], Guanliang Meng[1], Yuntao Tan[1], Jiang Hu[1], Xiaokang Li, Qihua Lai, Lingling Han[1], Depeng Wang[1], Fengyi Hu[2], Wen Wang[3,4,*], Shanlin Liu[1,5*]

1. GrandOmics Biosciences, Beijing, 102200, China

2. State Key laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Research Center for Perennial Rice Engineering and Technology of Yunnan, School of Agriculture, Yunnan University, Kunming, Yunnan, 650091, China

3. State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 650223 Kunming, Yunnan, China.

4. Center for Ecological and Environmental Sciences, Key Laboratory for Space Bioscience & Biotechnology, Northwestern Polytechnical University, 710072 Xi'an, China.

5. Department of Entomology, College of Plant Protection, China Agricultural University, 100193 Beijing, China

[#]Contribute equally

*Correspondence to Shanlin Liu: liushanlin@grandomics.com & Wen Wang: wwang@mail.kiz.ac.cn

Abstract

The availability of reference genomes has revolutionized the study of biology. Multiple competing technologies have been developed to improve the quality and robustness of genome assemblies during the last decade. The two widely-used long-read sequencing providers – Pacbio (PB) and Oxford Nanopore Technologies (ONT) – have recently updated their platforms: PB enables high throughput HiFi reads with base-level resolution with >99% and ONT generated reads as long as 2 Mb. We applied the two up-to-date platforms to one single rice individual and then compared the two assemblies to investigate the advantages and limitations of each. The results showed that ONT

ultralong reads delivered higher contiguity producing a total of 18 contigs of which ten were assembled into a single chromosome compared to that of 394 contigs and three chromosome-level contigs for the PB assembly. The ONT ultralong reads also prevented assembly errors caused by long repetitive regions for which we observed a total of 44 genes of false redundancies and ten genes of false losses in the PB assembly leading to over/under-estimation of the gene families in those long repetitive regions. We also noted that the PB HiFi reads generated assemblies with considerably fewer errors at the level of single nucleotide and small InDels than that of the ONT assembly which generated an average 1.06 errors per Kb and finally engendered 1,475 incorrect gene annotations via altered or truncated protein predictions.

Main text

The availability of reference genomes has revolutionized the study of biology – many diseases found their causative alleles thanks to the high quality human reference genome [1, 2]; the genomes of agricultural crops have tremendously accelerated our understanding on how artificial selection shaped plant traits and how, in turn, these plant traits may influence species interactions, e.g. phytophagous insects, in agriculture [3, 4]. During the last decade, multiple competing technologies have been developed to improve the quality and robustness of genome assemblies [5-8], enabling genome reference collecting of the tree of life [9-11]. To date, a large number of genomes have been assembled by Third Generation Sequencing (TGS) technologies which can produce individual reads in the range of 10~100 kbs or even longer [12-15]. Although the long-read still has a high error rate, it has been improving owing to the advances in sequencing chemistry and computational tools, e.g. Pacbio (PB) Single-molecule real-time (SMRT) sequencing platform released the Sequel II system of which the updated SMRT cell enabled high throughput HiFi reads using the circular consensus sequencing (CCS) mode to provide base-level resolution with >99% single-molecule read accuracy [16]; while the Oxford Nanopore Technologies (ONT) launched its PromethION

platform which can yield > 7 Tb per run and its ultralong sequencing application facilitates the achievement of complete genome - Telomere to Telomere (T2T) - by resolving long and complex repetitive regions for various species including *Homo sapien* [17]. Plenty of species begin to leverage the two cutting-edge sequencing systems; however, almost all chose one single sequencing system, either the PB or the ONT platform, to obtain their reference genomes [15, 18, 19]. Here we present one rice individual (variety 9311) that was sequenced and assembled independently using the two up-to-date systems, and then we compared the two assemblies to investigate the advantages and limitations of each.

Following DNA extraction from the rice sample, we sequenced the two extracts using ONT PromethION and PB Sequel II platforms, respectively. The PromethION generated a total of 92 Gb data with N50 of 41,473 bp and the Sequel II produced a total of 253 Gb data with each molecular fragment being sequenced 14.72 times on average and of an average length of 11,487 bp. We applied multiple software, including Canu1.9 [20], NextDenovo2.0-beta.1 (https://github.com/Nextomics/NextDenovo) and WTDBG2 [21] to assemble the rice genome for the ONT and PB dataset respectively. We selected the optimal assembly for each sequencing platform based on contig N50 and then got both polished (detailed in Supplementary Methods). The ONT assembly showed higher contiguity with a contig number of 18 and an N50 value of ca. 32 Mb in comparison to a contig number of 394 and N50 of 17 Mb for the PB assembly (Figure 1a & Table S1). Ten and three out of the total 12 autosomes were assembled into a single contig in the ONT and PB assembly, respectively. We identified telomers and centromeres for both assemblies (Supplementary Methods) and found that seven of them reached a T2T level assembly with no gaps and no Ns in between (Table S2). A genome completeness assessment using BUSCOv3.1.0 [22] finds both assemblies performed well with the ONT having a tiny improvement (98.62% vs 98.33%, Table S3). We mapped both assemblies to a high-quality rice (R498) genome reference [23] using Minimap2 [24]. Both assemblies showed good collinearity (Figure S1) and the

PB assembly contained more gaps in each chromosome compared to that of ONT (Figure 1a).

We then took Chr. 6 where ONT's one single contig (32,367,127 bp) corresponded to nine contigs (32,476,323 bp) of the PB assembly to investigate and visualize the incongruencies between them. For the nine contigs of PB assembled for the Chr. 6, four reached a length ≥ 6 Mb and five had a length of merely 10-70 kb. We investigated the three gaps where the top four PB contigs (named as PB-L1, PB-L2, PB-L3 and PB-L4 from 5' to 3'end, respectively) failed to connect (Figure 1b). We mapped the ONT ultralong reads to those gaps and confirmed their correctness through manual inspections by IGV plot [25](Figure S2). The gap #1 between PB-L1 and PB-L2 reached a length of 74,888 bp. One of the short PB contigs (PB-S1, length of 70,208 bp) had an overlap of ~10 kb with the 3'end of PB-L1, thus left the gap #1 a region of 15,722 bp that PB failed to cover (Figure 1c). We further examined the sequences obtained by ONT in and flanking this gap. It showed 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. The gap #2 between PB-L2 and PB-L3 characterized a region spanning up to 48 kb on the ONT assembly and is flanked by two tandem repeats of 14 kb in length. It was gone through by multiple ONT long reads (Figure S2), so can be successfully connected by the ONT assembly. The last gap between PB-L3 and PB-L4 can be connected by one short PB contig (PB-S2, 25,292 bp), which had 9,469 and 2,621 bp overlaps with 3'end of PB-L3 and 5'end of PB-L4, respectively. And it showed the same case as the gap #2, containing three tandem duplicates of length 23 kb that failed to be connected by PB HiFi reads. We found a total of 107 kb redundancies and 15 kb gaps on Chr. 6 owing to PB's incorrect assemblies, which corresponded to an excess of 13 annotated genes (Figure 2, Table S4). The genome-wide misassembled regions accumulated to a length of ~ 668 kb (534 kb redundancies and 134 kb gaps), hosting 54 annotated genes (44

redundancies and 10 loss, Table S4). As PB assembly did not generate any single contigs that ONT broke into multiple segments, we cannot find a counter case for comparison.

In addition to those gaps that PB failed to connect, we noticed that there were a bunch of small-scale mismatches between the two assemblies. Firstly, we extracted the reciprocal matches ⩾1 M between the two assemblies for comparison using QUAST 5.0.2 [26]. Then, we mapped the PB HiFi reads to both genome assemblies to filter out the innate discrepancies derived from diploid heterozygous states under the assumption that HiFi reads provide high level single base accuracy (Supplementary Methods). It showed that the ONT assembly, although polished using 70x Illumina's shotgun reads, still contained a large number of small-scale mis-assemblies. In total, we found 210,993 single nucleotide errors and 211,517 InDels (Mean: 1.39 bp, Figure S4) accounting for an average number of 1.06 errors per kb. However, instead of scattering evenly on the assembly, those errors formed into clusters (Figure S5). A further investigation for those regions showed ~ 94% of them have a shotgun read coverage ≤ 5, which explains why the last polishing step failed to fix those errors (Figure S6). About 7.48 % of those errors located on exons and affected ~ 2,415 exons (1,475 genes) to translate correctly to amino acid sequences. We did note that, however, the errors of HiFi reads may be enriched in sequences with particular characteristics, rather than completely random, for example, regions like simple sequence repeats and long homopolymers (Supplementary Methods, Figure S7) which may exacerbate the above error statistics for the ONT assembly.

In conclusion, our study investigated genome assembly qualities between the two up-to-date competing long read sequencing techniques - the PB's HiFi reads and the ONT's ultralong reads. It showed both techniques had their own merits with: (1) ONT ultralong reads delivered higher contiguity and prevented false redundancies caused by long repeats, which, in our case of rice genome, assembled 10 out of the 12 autosomes into one single contig, and (2) PB HiFi reads produced fewer errors at the level of single

nucleotide and small InDels and obtained more than 1,400 genes that incorrectly annotated in the ONT assembly due to its error prone reads. Therefore, we suggest that further genomic studies, especially genome reference constructions, should leverage both techniques to lessen the impact of assembly errors and subsequent annotation mistakes rooted in each. There is also an urgent demand for improved assembly and error correction algorithms to fulfill this task.

Availability of data and materials

We have all the data including two genome assemblies and their corresponding raw reads deposited on NCBI under the project ID PRJNA600693.

Competing interests

The authors declare that they have no competing interests.

References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: **Phenotypic impact of genomic structural variation: insights from and for human disease**. *NAT REV GENET* 2013, **14**(2):125.

2. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, Tanaka H, Taniguchi H, Kawakami Y, Ueno M: **Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer**. *NAT GENET* 2016, **48**(5):500.

3. Saxena RK, Edwards D, Varshney RK: **Structural variations in plant genomes**. *BRIEF FUNCT GENOMICS* 2014, **13**(4):296-307.

4. Chen YH, Gols R, Benrey B: **Crop domestication and its impact on naturally selected trophic interactions**. *ANNU REV ENTOMOL* 2015, **60**:35-58.

5. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth GT: **The complete genome of an individual by massively parallel DNA sequencing**. *NATURE* 2008, **452**(7189):872.

6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR: **Accurate whole human genome sequencing using reversible terminator chemistry**. *NATURE* 2008, **456**(7218):53.

7. Pushkarev D, Neff NF, Quake SR: **Single-molecule sequencing of an individual human genome**. *NAT BIOTECHNOL* 2009, **27**(9):847.

8. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M: **An integrated semiconductor device enabling non-optical genome sequencing**. *NATURE* 2011, **475**(7356):348.

9. Seberg O, Droege G, Barker K, Coddington JA, Funk V, Gostel M, Petersen G, Smith PP: **Global Genome Biodiversity Network: saving a blueprint of the Tree of Life – a botanical perspective**. *ANN BOT-LONDON* 2016, **118**(3):393-399.

10. Mukherjee S, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Göker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D: **1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life**. *NAT BIOTECHNOL* 2017, **35**(7):676.

11. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP: **Earth BioGenome Project: Sequencing life for the future of life**. *Proceedings of the National Academy of Sciences* 2018, **115**(17):4325-4333.

12. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M: **Resolving the complexity of the human genome using single-molecule sequencing**. *NATURE* 2015, **517**(7536):608.

13. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Spittle K, Hall R, Gu J, Lyons E: **Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum**. *NATURE* 2015, **527**(7579):508.

14. Gordon D, Huddleston J, Chaisson MJ, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW: **Long-read sequence assembly of the gorilla genome**. *SCIENCE* 2016, **352**(6281):e344.

15. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin C: **Improved maize reference genome with single-molecule technologies**. *NATURE* 2017, **546**(7659):524.

16. Wenger AM, Peluso P, Rowell WJ, Chang P, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND: **Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome**. *NAT BIOTECHNOL* 2019, **37**(10):1155-1162.

17. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA: **Telomere-to-telomere assembly of a complete human X chromosome**. *BioRxiv* 2019:735928.

18. Loman NJ, Quick J, Simpson JT: **A complete bacterial genome assembled de novo using only nanopore sequencing data**. *NAT METHODS* 2015, **12**(8):733.

19. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT: **Nanopore sequencing and assembly of a human genome with ultra-long reads**. *NAT BIOTECHNOL* 2018, **36**(4):338.

20. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation**. *GENOME RES* 2017, **27**(5):722-736.

21. Ruan J, Li H: **Fast and accurate long-read assembly with wtdbg2**. *BioRxiv* 2019:530972.

22. Seppey M, Manni M, Zdobnov EM: **BUSCO: Assessing Genome Assembly and Annotation Completeness**. In.: Springer; 2019: 227-245.

23. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, Ma B, Qi M, Li Y, Zhao X: **Sequencing and de novo assembly of a near complete indica rice genome**. *NAT COMMUN* 2017, **8**:15324.

24. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *BIOINFORMATICS* 2018, **34**(18):3094-3100.

25. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *NAT BIOTECHNOL* 2011, **29**(1):24.

26. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A: **Versatile genome assembly evaluation with QUAST-LG**. *BIOINFORMATICS* 2018, **34**(13):i142-i150.
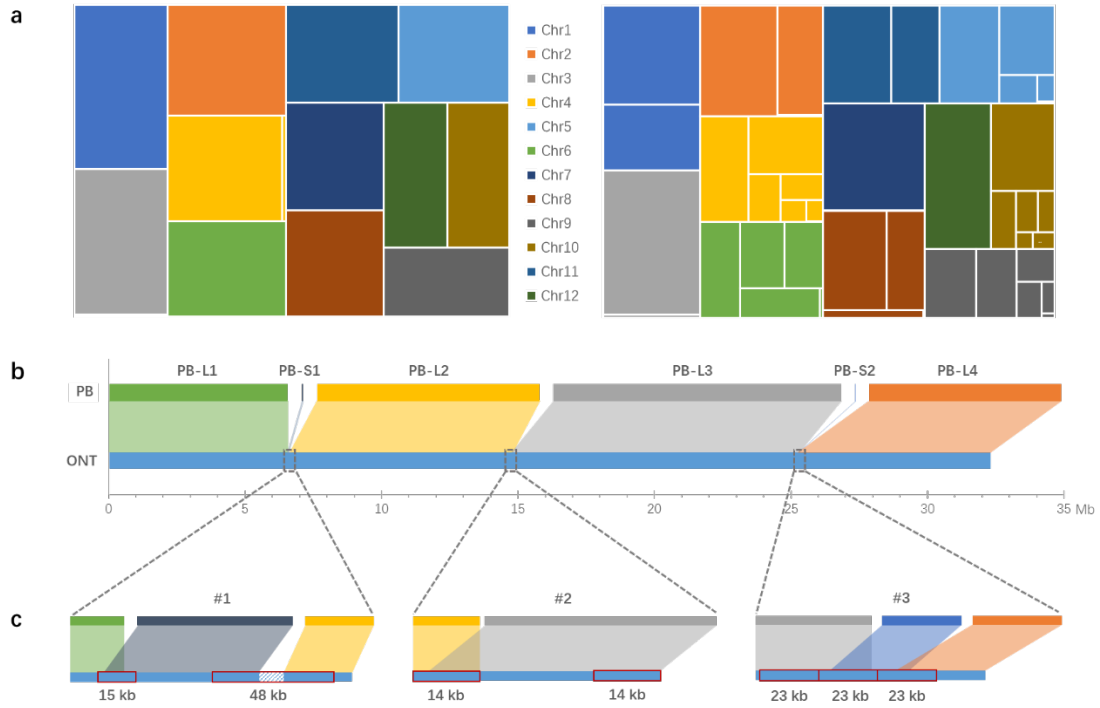
Figure 1. Contiguity of the ONT and PB assemblies. (a) Treemaps for contig length difference between the ONT (left) and PB (right) assembly; (b) The six PB contigs mapped to one ONT contig corresponding to Chr. 6; (c) Details of the three PB gaps. Red rectangles noted the repeat elements.
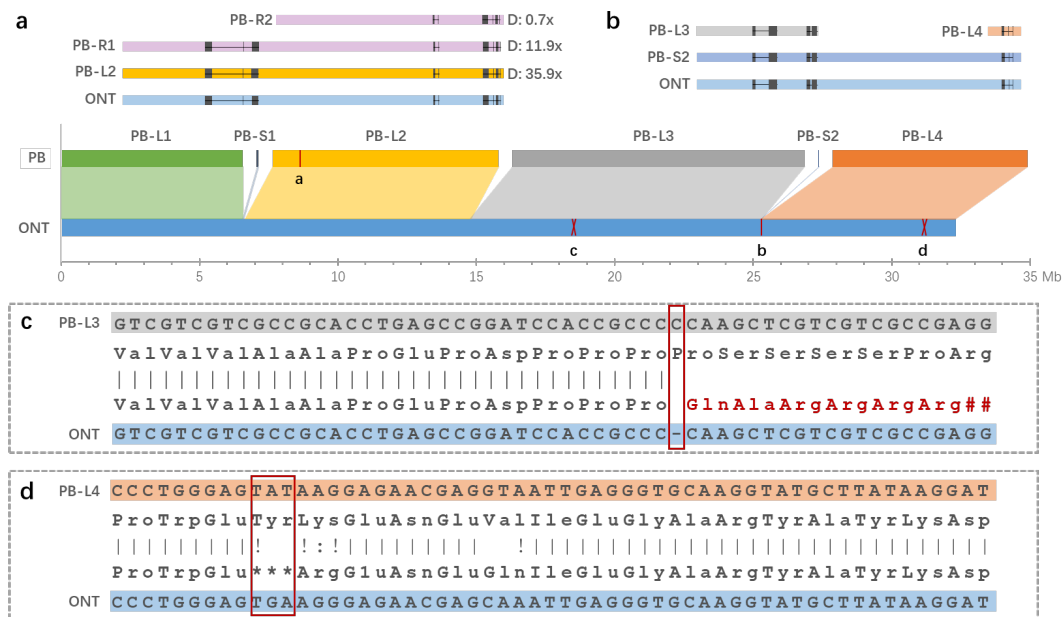
Figure 2. Assembly errors in which genes can be annotated. (a) An example shows gene gains that caused by assembly redundancies, of which the PB-R1 and PB-R2 had a similarity level of 99.67% and 99.51%, respectively, compared to the corresponding region on PB-L2, and "D" abbreviates from depth; (b) The gene redundancies caused by gaps that failed to be correctly connected by the PB assembly; (c) An example shows a 1-base deletion led to frameshift mistake for protein translation; (d) An example shows single base error led to stop codon gain and truncated protein translation.

9

Manuscript

Click here to access/download
Supplementary Material
Supplementary information.pdf