Au *Genome Biology* (2022) 23:21
https://doi.org/10.1186/s13059-022-02604-2

Genome Biology

EDITORIAL Open Access

The blooming of long-read sequencing reforms biomedical research

Kin Fai Au[1,2]

Correspondence: kinfai.au@osumc.edu
[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA
[2]Biomedical Informatics Shared Resources, The Ohio State University, Columbus, OH 43210, USA

Compared to Next Generation Sequencing (NGS), PacBio and nanopore sequencing offer ultra-long reads for single DNA/RNA molecules. These long reads are very informative to address omics problems where large-range complexity is involved, such as genome assembly, haplotyping, complex variant calling, and gene isoform identification. The single-molecule feature of long-read sequencing also allows for simultaneous measurements of base modifications together with other omics features, such as genomics and transcriptomics. This gives us unprecedented views on biomedical problems that have, until now, remained poorly characterized [1, 2]. Moreover, the accuracy, accessibility, and cost efficiency of long-read sequencing are improving dramatically, which boosts the long read-based research in many topics.

Therefore, we are not only in the midst of a new revolution in sequencing technology but also the next revolution in biomedical research. To timely and fully utilize the unique benefits of this technological breakthrough, there is much enthusiasm to develop new experimental and computational methods and apply long-read sequencing to diverse biomedical contexts. This special issue collects the latest work of several typical types of long read-based research in genomics, transcriptomics, and cancer diagnosis. Some of them aim to improve the existing analyses that rely on the NGS-based methods, and some others are unique applications for long-read sequencing, such as nanopore adaptive sampling.

Genome assembly is one of the earliest and the most popular applications of long reads. As they cover many single nucleotide polymorphisms (SNPs), long reads are useful to advance genome assembly to the haplotype resolution. The new software phasebook adapts a divide-and-conquer strategy to improve the coverage of haplotype-resolved de novo genome assembly [3]. Since sample preparation, such as high-molecular-weight DNA extraction, could influence the data quality and thus assembly significantly, the end-to-end plant genome assembly workflow LeafGO optimizes the steps from sample preparation to computational analysis [4]. In addition to the application for the diploid genomes, LeafGO was also tested in the allotetraploid genome of *Arachis hypogaea*. Improved haplotype-resolved assembly is very beneficial to many research and applications, such as precision medicine and evolutionary biology. For example, Xue et al. used PacBio HiFi reads to create a high-quality and nearly gap-free

diploid genome of zig-zag eel so that they could perform a high-resolution comparison of the homomorphic pair of the sex chromosomes to investigate their recombination and differentiation [5]. The power of long reads was also shown in characterizing structural variations and repetitive elements [6, 7].

In the field of transcriptomics, abundance estimation is the basis of many other analyses, so Hu et al. published LIQA, a more sophisticated method for gene isoform quantification by long reads other than simply using read counts as the expression index in the previous studies [8]. In parallel, the interests of single-cell sequencing are emerging in the field of long read-based transcriptomics research. In particular, Tian et al. optimized a droplet-based protocol for generating high-quality single-cell sequencing data of both short reads and long reads and also established a bioinformatics pipeline FLAMES for comprehensive analyses, such as isoform identification and mutation detection in single cells [9]. Considering cost efficiency, Rebboah et al. reported a protocol LR-Split-seq that integrates the combinatorial barcoding of Split-seq with long-read sequencing to achieve differential gene isoform expression analysis at the single-cell level [10]. The application of LR-Split-seq to the C2C12 myogenic system found the distinct patterns of alternative transcription start sites and/or alternative internal exon usage in different cell clusters. Besides cDNA sequencing, nanopore sequencing can measure native RNA molecules directly, so Schulz et al. were able to revisit the reliability of exitron identification by comparing the data of direct RNA sequencing and cDNA sequencing [11]. They found that dozens of exitrons may be artifacts of reverse transcription, highlighting the value and importance of validation by direct RNA sequencing.

To leverage the time-/cost-efficiency of nanopore sequencing for clinical usage, Thirunavukarasu et al. developed a cancer screening protocol "Oncogene Concatenated Enriched Amplicon Nanopore Sequencing (OCEANS)" targeting the somatic mutations with low variant allele frequency [12]. They demonstrated the accuracy by applying the specific panels of recurrent mutations to four cancer types and showed it a possible measure for rapid and affordable clinical sequencing.

A few new efforts are specific for long-read sequencing. For example, adaptive sampling is a unique application of nanopore sequencing to enrich target elements of interest—real-time analysis of the raw electrical signals determines whether the molecules are ejected, or the data collection continues. Bao et al. developed the first deep learning-based software SquiggleNet to improve the analysis speed and computing memory usage [13]. Martin et al. established a mathematical model to evaluate how a set of factors, such as molecule length, influence the enrichment performance, so that the output can be predicted and a guideline of adaptive sampling was also provided [14].

Considering the rapid growth of computational methods, experimental techniques, and applications, benchmarking is a critical type of effort to optimize and promote the usage of the long-read sequencing, especially for many starters with limited experience. For instance, Chen et al. developed a computational platform Inspector to evaluate genome assembly [15]. Because of the large variance of long-read sequencing data quality, such as read length and error profile, the performance of assemblers was examined in different data scenarios (e.g., PacBio CLR and HiFi reads and nanopore data). It is indeed a good practice to benchmark long read-based methods, and doing so can provide

more specific guidelines for data collection and software selection. Liu et al. completed a comprehensive survey for nanopore sequencing-based 5mC detection across different genomic contexts, CpG site coverage, and computational resources [16]. In addition to the single-site resolution, the "per-read" accuracy, i.e., detection at the single-molecule level, was also tested, which is a new view for advancing epigenetics research. Like the era of NGS, consortium-scale efforts of method benchmarking and construction of omics landscapes will be very beneficial to the community of long-read sequencing by providing useful analysis guidelines and valuable data resources. For instance, the Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium is now organizing a large-scale survey of different protocols and software for long read-based RNA-seq [17].

Although it is not possible to include all significant research of long-read sequencing within a single special issue, this collection of articles represents the emerging interests and trends of long read-based method development and applications. We foresee much more creative and impactful research of long-read sequencing in the coming years.

Acknowledgements
K.F.A. is grateful for the support from an institutional fund from the Department of Biomedical Informatics, The Ohio State University, and the National Institutes of Health (R01HG008759, R01HG011469, and R01GM136886). K.F.A. would like to thank Dr. Barbara Cheifet for critically reading and editing the manuscript.

Author's contributions
The author read and approved the final manuscript.

Declarations

Competing interests
The author declares that he has no competing interests.

Published online: 12 January 2022

References

1. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteome Bioinforma. 2015;13(5):278–89. https://doi.org/10.1016/j.gpb.2015.08.002.
2. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol. 2021;39(11):1348–65. https://doi.org/10.1038/s41587-021-01108-x.
3. Luo X, Kang X, Schonhuth A. Phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. Genome Biol. 2021;22(1):299. https://doi.org/10.1186/s13059-021-02512-x.
4. Driguez P, Bougouffa S, Carty K, Putra A, Jabbari K, Reddy M, et al. LeafGo: leaf to genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. Genome Biol. 2021;22(1):256. https://doi.org/10.1186/s13059-021-02475-z.
5. Xue L, Gao Y, Wu M, Tian T, Fan H, Huang Y, et al. Telomere-to-telomere assembly of a fish Y chromosome reveals the origin of a young sex chromosome pair. Genome Biol. 2021;22(1):203. https://doi.org/10.1186/s13059-021-02430-y.
6. Quan C, Li Y, Liu X, Wang Y, Ping J, Lu Y, et al. Characterization of structural variation in Tibetans reveals new evidence of high-altitude adaptation and introgression. Genome Biol. 2021;22(1):159. https://doi.org/10.1186/s13059-021-02382-3.
7. Chiu R, Rajan-Babu IS, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021;22(1):224. https://doi.org/10.1186/s13059-021-02447-3.
8. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K. LIQA: long-read isoform quantification and analysis. Genome Biol. 2021;22(1):182. https://doi.org/10.1186/s13059-021-02399-8.
9. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22(1):310. https://doi.org/10.1186/s13059-021-02525-6.
10. Rebboah E, Reese F, Williams K, Balderrama-Gutierrez G, McGill C, Trout D, et al. Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq. Genome Biol. 2021;22(1):286. https://doi.org/10.1186/s13059-021-02505-w.
11. Schulz L, Torres-Diz M, Cortes-Lopez M, Hayer KE, Asnani M, Tasian SK, et al. Direct long-read RNA sequencing identifies a subset of questionable exitrons likely arising from reverse transcription artifacts. Genome Biol. 2021;22(1):190. https://doi.org/10.1186/s13059-021-02411-1.
12. Thirunavukarasu D, Cheng LY, Song P, Chen SX, Borad MJ, Kwong L, et al. Oncogene concatenated enriched amplicon nanopore sequencing for rapid, accurate, and affordable somatic mutation detection. Genome Biol. 2021;22(1):227. https://doi.org/10.1186/s13059-021-02449-1.
13. Bao Y, Wadden J, Erb-Downward JR, Ranjan P, Zhou W, McDonald TL, et al. SquiggleNet: real-time, direct classification of nanopore signals. Genome Biol. 2021;22(1):298. https://doi.org/10.1186/s13059-021-02511-y.

14. Martin S, Heavens D, Lan Y, Horsfield S, Clark MD, Leggett RM. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. bioRxiv. 2021.
15. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z. Accurate long-read de novo assembly evaluation with Inspector. Genome Biol. 2021;22(1):312. https://doi.org/10.1186/s13059-021-02527-4.
16. Liu Y, Rosikiewicz W, Pan Z, Jillette N, Wang P, Taghbalout A, et al. DNA methylation-calling tools for Oxford Nanopore sequencing: a survey and human epigenome-wide evaluation. Genome Biol. 2021;22(1):295. https://doi.org/10.1186/s13059-021-02510-z.
17. Pardo-Palacios F, Reese F, Carbonell-Sala S, Diekhans M, Liang C, Wang D, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Res Square. 2021.

Publisher's Note