JIG Journal of Interdisciplinary Genomics
https://doi.org/10.22742/JIG.2024.6.2.15
Journal of Interdisciplinary Genomics 2024;6(2):15-20
eISSN 2671-6771

The Growing Potential of Long-Read Sequencing in Identifying Previously Elusive Causative Variants in Patients with Undiagnosed Rare Diseases

Yeonsong Choi[1,2], David Whee-Young Choi[1,2], Hyeyeon Won[1,2], Semin Lee[1,2]

[1]Department of Biomedical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Korea
[2]Korean Genomics Center, UNIST, Ulsan, Korea

Rare diseases, largely driven by genetic factors, present significant diagnostic challenges due to their complex genomic variations. Traditional short-read sequencing methods, such as whole-exome sequencing and whole-genome sequencing, are widely used to detect genomic alterations in a time- and cost-effective manner. However, some rare conditions are often left undiagnosed due to the technical limitations of current sequencing platforms. To overcome these limitations, long-read sequencing (LRS) technology has been applied to various fields of clinical research including rare diseases. With LRS, researchers are able to accurately characterize complex variants such as structural variations, tandem repeats, transposable elements, and transcript isoforms. This review article explores the current applications of LRS in rare disease research, highlighting its potential in identifying previously elusive causative variants in undiagnosed rare diseases.

Key words: Long-read sequencing, Rare diseases, Mendelian disorder, Structural variations, Tandem repeats, Transposable elements, Transcript isoforms

REVIEW ARTICLE

Received: August 29, 2024
Revised: September 5, 2024
Accepted: September 12, 2024

Correspondence to: Semin Lee, PhD
Department of Biomedical Engineering, UNIST, Ulsan 44919, Korea
Tel: +82+52-217-2663
Fax: +82+52-217-3229
E-mail: seminlee@unist.ac.kr

ORCID
https://orcid.org/0000-0002-9015-6046

INTRODUCTION

Rare diseases, typically defined as conditions affecting fewer than 1 in 2,000 individuals, represent a significant global health challenge. It is estimated that over 400 million people worldwide suffer from such conditions, with approximately 80% of these diseases believed to be genetic in origin [1,2]. The Online Mendelian Inheritance in Man (OMIM) database currently documents over 6,300 distinct disease-gene associations, highlighting the genetic diversity and complexity inherent in rare diseases.

In South Korea, rare diseases are defined more narrowly, where conditions affecting fewer than 20,000 individuals are considered rare [3]. Among these, diseases with fewer than 200 patients are classified as extremely rare. The limited number of patients, along with the requirement of specialized knowledge for diagnosis, often complicates the identification and management of these diseases. As of November 2023, approximately 1,250 diseases are officially recognized as rare in South Korea, where the national healthcare system provides coverage for the associated medical expenses. This underscores the significant burden that rare diseases place on patients and healthcare systems alike, given the persistent health risks and complications associated with these conditions.

Currently, the most widely used next-generation sequencing (NGS) technology for diagnosis of rare diseases is whole-exome sequencing (WES). This method targets the exonic region, which contains the protein-coding sequences of the genome. Although the exome only comprises 1%–2% of the genome, it contains 85% of genetic variants that have a high impact on the pathogenicity of diseases [4]. In addition, WES allows for a high sequencing depth in an accurate and cost-effective manner compared to other sequencing platforms, enabling the reliable detection of genetic variants such as single nucleotide variants (SNVs) and short insertions and deletions (INDELs). Moreover, data generated by WES is relatively small and manageable, greatly reducing computational burden and time required for analysis. These factors make WES meritable for researchers with large patient cohorts and even make trio analysis feasible, which is very important in identifying causative variants.

Despite the advantages described above, WES is limited to the exonic region; therefore, non-coding variants and structural variations (SVs) in the intronic regions cannot be detected. Hence, whole-genome sequencing (WGS) is used when genome-wide analysis is necessary. SNVs and INDELs in the non-coding regions in addition to SVs can be detected with WGS, which can make this sequencing platform more appropriate than WES depending on the aim of the study. However, short-read WGS produces sequence reads with an average length of 100 bp. As such, analysis of complex SVs or repeat regions with lengths over 100 bp is still challenging.

Diagnosis rates using WES or WGS data range from 25%–50%, yielding slightly better rates with WGS [5,6]. In other words, roughly 50% of patients remain undiagnosed. Among many factors behind the difficulties hindering diagnosis, the technical limitations of short-read sequencing (SRS) contribute considerably by limiting the effectiveness in detecting and analyzing complex genetic variations. To address these issues, developments in long-read sequencing (LRS) have emerged recently and attempts to identify previously undetected causative variants have increased using this technology.

As of date, the most prevalent LRS platforms have been developed by PacBio and Oxford Nanopore Technologies (ONT), where each has its distinctions. PacBio's LRS technology is known for its overall high data quality, as molecules can be sequenced multiple times to generate low-error data [7]. However, LRS data generation with PacBio has a higher cost and requires larger amounts of higher-quality DNA [8]. On the other hand, ONT's LRS technology provides a higher throughput at a lower cost, which is an important factor for improving the efficiency and scalability of research projects. Furthermore, longer mappable reads are achievable with ONT, but generally shorter reads are generated compared to PacBio [9]. As such, both technologies have their differences and researchers should choose LRS platforms that are the most appropriate depending on the objectives of their research.

The introduction of LRS technology has enabled the characterization of complex SVs, repeat regions, and phases that were previously difficult to identify, as well as the characterization of gene isoforms [10]. On the other hand, LRS also comes with certain limitations, including high cost and low accuracy per read. Furthermore, current bioinformatic analysis methods for LRS are limited and are still under development. Nevertheless, LRS definitely has its strengths, and various efforts are continuously being made to minimize error and improve analysis techniques. With advancements in this technology, LRS has significant potential to help characterize the complex genomic changes occurring in rare diseases and accurately diagnose affected patients.

This review article aims to introduce LRS technology and outline the applications of LRS in identifying causative variants of mendelian disorders or rare diseases.

COMPLEX STRUCTURAL VARIATIONS

Structural variations (SVs), which are large genomic alterations with sizes ranging from 50 bp to over 1 kb, are commonly associated with genetic diseases [11]. These alterations include inversions, translocations, insertions, deletions and can have a significant impact on a gene's function. Due to their sizes, accurately depicting large SVs with SRS serves to be quite difficult. Furthermore, current NGS techniques have limited capabilities in detecting SVs in regions with complex or repeated sequences. To address this problem, researchers have adopted LRS to characterize these alterations effectively (Fig. 1A). With LRS, it is possible to survey longer portions of the genome and in turn, observe SVs with higher accuracy [12]. Merker et al. [13] performed low-coverage genome LRS to identify SVs that could not be detected using SRS in a patient with multiple neoplasia and cardiac myxomata. As a result, they identified over 6,000 insertions and deletions larger than 50 bp, including a pathogenic 2,184 bp deletion overlapping with an exon of the *PRKAR1A* gene, which is associated with autosomal dominant Carney complex. Similar studies have been performed using PacBio LRS as well as Nanopore LRS techniques. In a study by Damián et al. [14], LRS was performed on two patients with congenital aniridia for whom causative variants were not identified through SRS analysis. This approach revealed pathogenic SVs in the *PAX6* gene responsible for con-

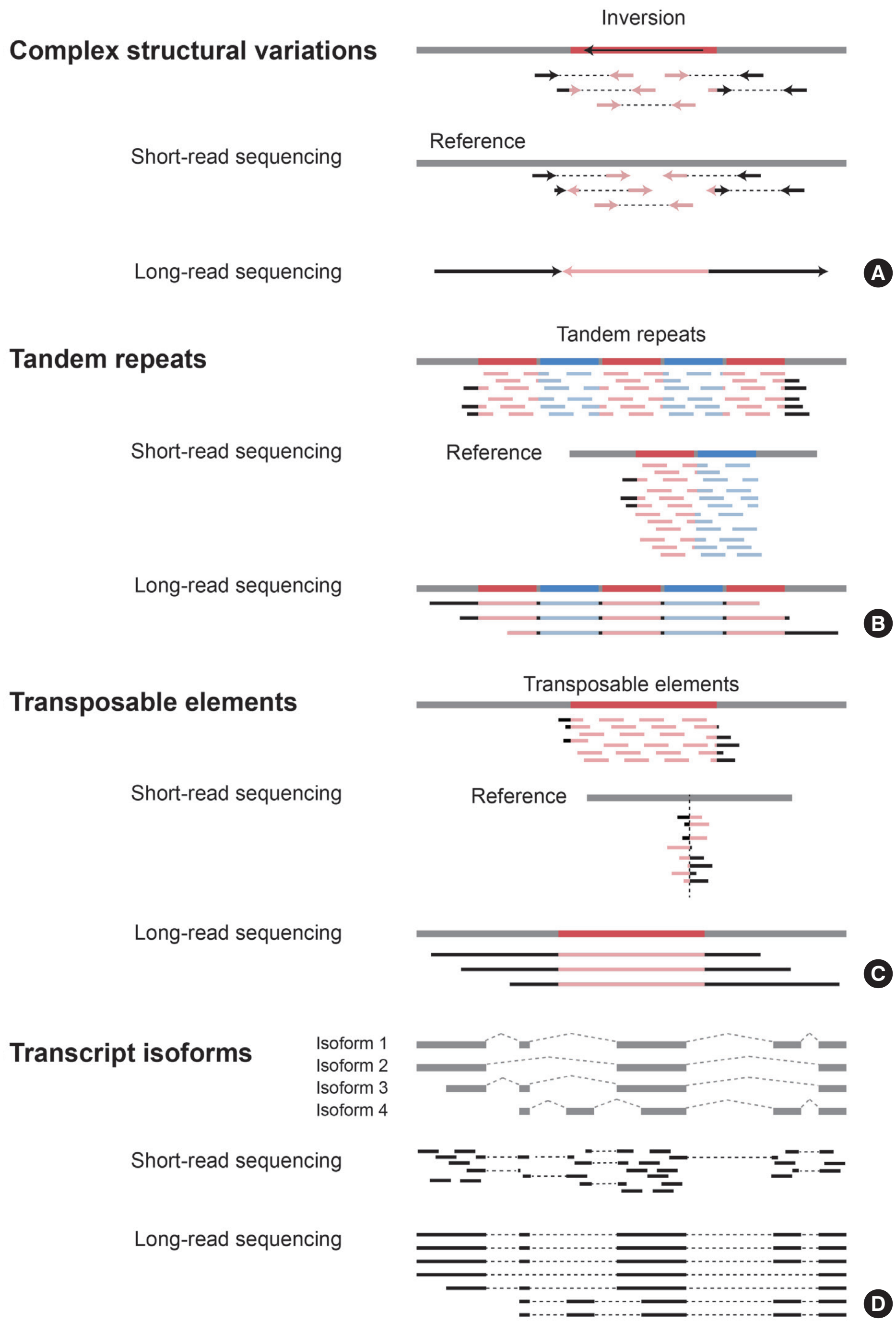

Fig. 1. Identifying causative variants using short-read sequencing vs. long-read sequencing. (A) Detection of complex structural variations. Diagram showing an example of sequencing complex structural variations. The black and light red arrows indicate reads mapped to the complex structural variation. (B) Detection of tandem repeats. The colored regions indicate repeats occurring in the genome and different colors were used to differentiate each repeat unit. The light red and light blue lines show reads mapped to the repeat region. (C) Detection of transposable elements. The red regions indicate transposable elements. The light red lines show reads mapped to the transposable element region. (D) Detection of transcript isoforms. The black regions are reads mapped to the transcripts.

genital aniridia, including a 4.9 Mb inversion in intron 7 and a t(6;11) balanced translocation. In another study, long-read WGS was performed on a patient suspected of having autosomal recessive glycogen storage disease type Ia (GSD-Ia) in whom the causative variant had not been properly identified by WES. While WES had detected only a single heterozygous pathogenic variant in the *G6PC* gene, LRS revealed the presence of a 7.1 kb deletion covering two exons on the other allele (Miao et al.) [15]. Furthermore, there are studies where LRS has been performed on a larger cohort as opposed to a few samples. Miller et al. conducted targeted LRS on 30 patients with previously identified causative variants and 10 patients without an accurate diagnosis. They successfully confirmed the previously identified pathogenic SVs and further detected pathogenic/likely pathogenic variants in 6 out of the 10 undiagnosed patients [16]. These studies demonstrate the potential of LRS in characterizing complex SVs, especially those that could not be identified with conventional SRS methods.

TANDEM REPEATS

Tandem repeats, which are short copies of DNA sequences that are repeated multiple times in the genome, are highly polymorphic and are an important source of genetic variation [17]. The size of these repeated sequences range from a few base pairs to hundreds of base pairs. In the case of SRS, the read length is approximately 100 bp, rendering large tandem repeats difficult to be mapped accurately to the genome. On the other hand, LRS allows the detection of repeated sequences in large regions, which makes it suitable for tandem repeat analysis (Fig. 1B). In a study by Mizuguchi et al. [18], LRS was used to precisely investigate the causative variants in a benign adult familial myoclonus epilepsy (BAFME) family. As a result, they found a 4,661 bp heterozygous repeat insertion in the *SAMD12* intron region. Also, LRS from ONT was used to identify repeat expansions. With this approach, Sone et al. [19] confirmed that the 5' GCC repeat expansion in *NOTCH2NLC* was found only in neuronal intranuclear inclusion disease family members. These findings highlight the advantages of using LRS technology to provide deeper insight into complex diseases through accurate detection of tandem repeats and repeat expansions.

TRANSPOSABLE ELEMENTS

Transposable elements (TEs) have also been reported to play a role in rare diseases [20]. TEs are also known as "jumping genes," and as the name suggests, they are mobile DNA sequences that can move to different genomic locations. They can be found throughout all living organisms and are another source of evolution and genome reorganization. TE integration in the protein-coding region can induce gene dysfunction, while integration in the intronic region can induce alternative splicing events and in turn, abnormal gene expression [21]. A constraint of short-read WGS is that the generated reads are too short to fully cover each TE copy. Thus, only single nucleotide polymorphisms within a TE, or partial reads containing TE and genome junctions, can be mapped. Conversely, LRS overcomes this obstacle by generating reads of sufficient length to identify TE insertions (Fig. 1C) [22]. Zhou et al. [23] identified 90 L1Hs insertions in human cell lines that were not detected by previous SRS studies by developing computational software (PALMER) to detect LINE-1 insertions from PacBio reads. Moreover, LRS can be used in conjunction with other genomic analysis methods for TE analysis. Aneichyk et al. [24] also used multiple approaches, including PacBio SMRT in a X-Linked Dystonia-Parkinsonism (XDP) cohort to analyze the cause of the disease. They identified an SVA (SINE-VNTR-Alu) insertion in the *TAF1* gene that was exclusive to the XDP probands. In a different study, Fernández-Suárez et al. [25] employed LRS to detect an Alu retrotransposon insertion in the *EYS* gene from patients with retinitis pigmentosa, which was not discovered with their previous targeted SRS results. With LRS, researchers are able to discern genomic changes that span the genome in broad segments, which can contribute to fully characterizing presently ambiguous regions of the genome.

TRANSCRIPT ISOFORMS

Transcript isoforms, which are different versions of mRNA produced from the same gene, represent another factor contributing to the genetic alterations that can lead to rare diseases [26]. These mRNA sequences are produced as a result of alternative splicing events, which can cause the synthesis of abnormal transcripts or change the expression levels of otherwise normally produced genes. With SRS, identifying and analyzing the transcript isoforms coded by genes usually required the prediction of these various isoforms using average read depth. In contrast, LRS can sequence whole RNA molecules, and can detect novel isoforms, along with their expression levels (Fig. 1D). In a study by Stergachis et al. [27], they performed full-length long-read isoform sequencing to establish the consequence of a *MFN2* intron branch point variant. They confirmed that the variant produces five altered splicing transcripts that disrupt the open reading frames responsible for Charcot-Marie-Tooth disease, axonal, type 2A (CMT2A). This illustrates that LRS is not limited to DNA and can be applied to research that requires detection of transcriptomic isoforms.

CONCLUSION

In this review article, we introduced the ongoing research on the applications of LRS technology in effectively identifying genomic variations that previously could not be detected with conventional methods. While acknowledging the current limitations of LRS, it is important to recognize the advantages that are accompanied. With further improvements in the LRS platform and bioinformatic methodology, we believe that this technology has the potential to expand our knowledge on the complex mechanisms underlying rare diseases, ultimately aiming at improved patient diagnosis and appropriate treatment.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Ferreira CR. The burden of rare diseases. Am J Med Genet A 2019;179(6):885-92. doi: 10.1002/ajmg.a.61124.
2. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020;28(2):165-73. doi: 10.1038/s41431-019-0508-0.
3. Lim SS, Lee W, Kim YK, Kim J, Park JH, Park BR, et al. The cumulative incidence and trends of rare diseases in South Korea: a nationwide study of the administrative data from the National Health Insurance Service database from 2011-2015. Orphanet J Rare Dis 2019;14(1):49. doi: 10.1186/s13023-019-1032-6.
4. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet 2014;30(9):418-26. doi: 10.1016/j.tig.2014.07.001.
5. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The genetic basis of mendelian phenotypes: discoveries, challenges, and opportunities. Am J Hum Genet 2015;97(2):199-215. doi: 10.1016/j.ajhg.2015.06.009.
6. Sullivan JA, Schoch K, Spillmann RC, Shashi V. Exome/Genome sequencing in undiagnosed syndromes. Annu Rev Med 2023;74:489-502. doi: 10.1146/annurev-med-042921-110721.
7. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics 2024;116(3):110842. doi: 10.1016/j.ygeno.2024.110842.
8. Oehler JB, Wright H, Stark Z, Mallett AJ, Schmitz U. The application of long-read sequencing in clinical settings. Hum Genomics 2023;17(1):73. doi: 10.1186/s40246-023-00522-3.
9. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. Nat Rev Genet 2020;21(10):597-614. doi: 10.1038/s41576-020-0236-x.
10. Maestri S, Maturo MG, Cosentino E, Marcolungo L, Iadarola B, Fortunati E, et al. A long-read sequencing approach for direct haplotype phasing in clinical settings. Int J Mol Sci 2020;21(23):9177. doi: 10.3390/ijms21239177.
11. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet 2011;12(5):363-76. doi: 10.1038/nrg2958.
12. Yu SY, Xi YL, Xu FQ, Zhang J, Liu YS. Application of long read sequencing in rare diseases: the longer, the better? Eur J Med Genet 2023;66(12):104871. doi: 10.1016/j.ejmg.2023.104871.
13. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. Genet Med 2018;20(1):159-63. doi: 10.1038/gim.2017.86.
14. Damián A, Núñez-Moreno G, Jubin C, Tamayo A, de Alba MR, Villaverde C, et al. Long-read genome sequencing identifies cryptic structural variants in congenital aniridia cases. Hum Genomics 2023;17(1):45. doi: 10.1186/s40246-023-00490-8.
15. Miao H, Zhou J, Yang Q, Liang F, Wang D, Ma N, et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. Hereditas 2018;155:32. doi: 10.1186/s41065-018-0069-1.
16. Miller DE, Sulovari A, Wang T, Loucks H, Hoekzema K, Munson KM, et al. Targeted long-read sequencing identifies missing disease-causing variation. Am J Hum Genet 2021;108(8):1436-49. doi: 10.1016/j.ajhg.2021.06.006.
17. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 2018;19(5):286-98. doi: 10.1038/nrg.2017.115.
18. Mizuguchi T, Toyota T, Adachi H, Miyake N, Matsumoto N, Miyatake S. Detecting a long insertion variant in SAMD12 by SMRT sequencing: implications of long-read whole-genome sequencing for repeat expansion diseases. J Hum Genet 2019;64(3):191-7. doi: 10.1038/s10038-018-0551-7.
19. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. Nat Genet 2019;51(8):1215-21. doi: 10.1038/s41588-019-0459-y.
20. Chénais B. Transposable elements and human diseases: mechanisms and implication in the response to environmental pollutants. Int J Mol Sci 2022;23(5):2551. doi: 10.3390/ijms23052551.
21. Payer LM, Burns KH. Transposable elements in human genetic disease. Nat Rev Genet 2019;20(12):760-72. doi: 10.1038/s41576-019-0165-8.
22. Smits N, Faulkner GJ. Nanopore sequencing to identify transposable element insertions and their epigenetic modifications. Methods Mol Biol 2023;2607:151-71. doi: 10.1007/978-1-0716-2883-6_9.
23. Zhou W, Emery SB, Flasch DA, Wang Y, Kwan KY, Kidd JM, et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. Nucleic Acids Res 2020;48(3):1146-63. doi: 10.1093/nar/gkz1173.

24. Aneichyk T, Hendriks WT, Yadav R, Shin D, Gao D, Vaine CA, et al. Dissecting the Causal Mechanism of X-Linked dystonia-parkinsonism by integrating genome and transcriptome assembly. Cell 2018;172(5):897-909.e21. doi: 10.1016/j.cell.2018.02.011.
25. Fernández-Suárez E, González-Del Pozo M, Méndez-Vidal C, Martín-Sánchez M, Mena M, de la Morena-Barrio B, et al. Long-read sequencing improves the genetic diagnosis of retinitis pigmentosa by identifying an Alu retrotransposon insertion in the EYS gene. Mob DNA 2024;15(1):9. doi: 10.1186/s13100-024-00320-1.
26. Ergin S, Kherad N, Alagoz M. RNA sequencing and its applications in cancer and rare diseases. Mol Biol Rep 2022;49(3):2325-33. doi: 10.1007/s11033-021-06963-0.
27. Stergachis AB, Blue EE, Gillentine MA, Wang LK, Schwarze U, Cortés AS, et al. Full-length isoform sequencing for resolving the molecular basis of charcot-Marie-Tooth 2A. Neurol Genet 2023;9(5):e200090. doi: 10.1212/NXG.0000000000200090.