

Application of Whole Genome Long-read Sequencing Technology in Disease

Contents

1. Overview of Whole Genome Sequencing Technology

- 1.1 Human Whole Genome Sequencing and Long-read Sequencing Technologies.....1
- 1.2 Two Main Long-read Sequencing Technologies.....2

2. Applications of Long-read Whole Genome Sequencing Technology in Human Disease Diagnosis

- 2.1 Identifying Neurodegenerative Brain Diseases.....2
- 2.2 Detecting Genetic Diseases Caused by Genetic Variations.....4
- 2.3 Detecting Cancer.....5
- 2.4 Identifying the Methylation Status of Disease-related Genes.....6
- 2.5 Applying Third-generation Whole Genome Low-depth Sequencing to Disease Detection.....6

3. Conclusion

- 3.1 Advantages and Prospects of Long-read Whole Genome Sequencing Technology.....7

1. Overview of Whole Genome Sequencing Technology

1.1 Human Whole Genome Sequencing and Long-read Sequencing Technologies

Whole genome sequencing (WGS) techniques involve sequencing an entire genome or a set of genomes within a population, then performing bioinformatics analyses on the genomic data at either the individual or population level, respectively. WGS provides comprehensive information on many types of genomic variation such as **single nucleotide polymorphisms (SNPs)**, **insertions/deletions (Indels)**, **structural variants (SVs)**, and **copy number variations (CNVs)**. Using that detailed genomic data, researchers and clinicians can identify inherited



disorders, characterize mutations that drive cancer progression, and even track disease outbreaks. In the context of human health, WGS is a powerful tool for identifying and investigating diseases such as cancers, rare diseases, neurodegenerative diseases, and movement disorders, as well as for routine variant detection and genetic disease testing.

Short-read sequencing, especially **next-generation sequencing (NGS)** using Illumina machines and methods, has been an incredibly powerful tool of the genomic revolution. However, applying short-read technologies to SV detection and to genome assembly more broadly has revealed that their limited read length is a major shortcoming.^[1] Read lengths less than 300 bases are too short to detect more than 70% of human genome structural variation.^[2] In contrast, long-read or **“third-generation” sequencing (TGS)** technologies can generate continuous sequence reads directly from native DNA. These read lengths range from **10 kilobases (kb)** to **several megabases (mb)** in length, which, along with recent developments in throughput and accuracy, has substantially increased the utility and applicability of long-read technologies. The wealth of additional information afforded by single-molecule TGS, compared with short-read sequencing, promises a more comprehensive understanding of genetic, epigenetic, and transcriptomic variation and how those types of variation relate to human phenotypes.

TGS technologies are widely used in **early non-invasive diagnosis of tumors** and **genetic mutation detection in genetic diseases**. In addition to generating long read lengths, TGS does not have GC bias and it is advantageous for accurately sequencing highly repetitive and complex regions. TGS long reads provide greater coverage for samples with atypical clinical symptoms and for genetically complex tumors with multiple structural variations. They also improve disease detection rates, thereby compensating for the limitations of NGS technologies for detecting structural variation.

1.2 Two Main Long-read Sequencing Technologies

The two most widely used commercial long-read technologies are Single Molecule Real-time (SMRT) sequencing by Pacific Biosciences (PacBio) with an average read length of ~20 kb (>99.9% accuracy for HiFi reads) and nanopore sequencing by Oxford Nanopore Technologies (ONT) with an average read length of ~100 kb for ultra-long reads (~99% accuracy for R10.4). Their distinct sequencing principles and approaches to data generation yield sequencing reads with varied lengths, error rates, and throughputs. Researchers may find that one long-read sequencing technology meets their research goals and resource requirements better than the other, depending on the application. Both techniques are continually evolving.^[2]



PacBio Sequel II / II e



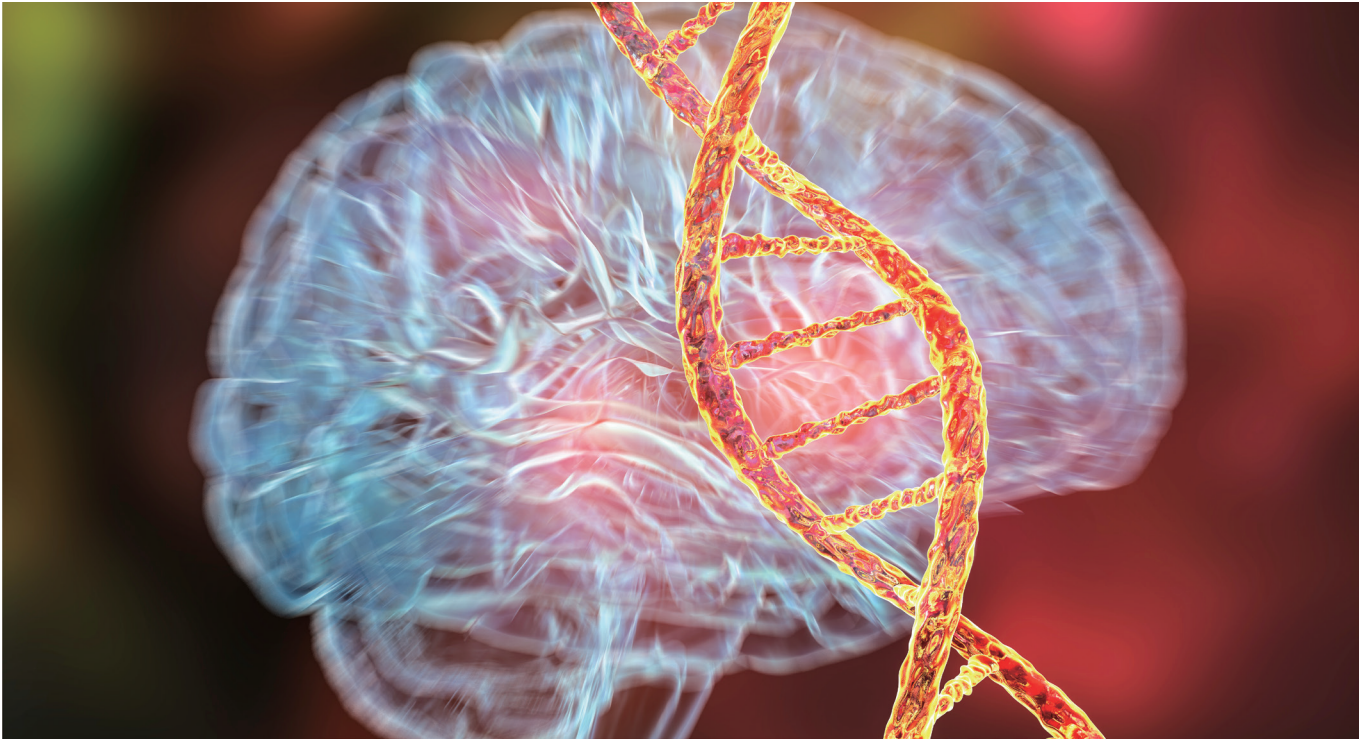
Nanopore PromethION

2. Applications of Long-read Whole Genome Sequencing Technology in Human Disease Diagnosis

TGS-based WGS strategies have diverse applications with respect to human disease. Long-read WGS can be used to identify and characterize a variety of genetic and structural genomic factors, including **mutations**, ***de novo* mutations (DNMs)**, **InDels**, and **complex SVs** that contribute to neurodegenerative brain diseases and cancers. They can also be used in conjunction with other technologies such as single-cell analysis and methylation status.

2.1 Identifying Neurodegenerative Brain Diseases

Neurodegenerative brain diseases are generally characterized by protein aggregates that accumulate in the brain and lead to progressive deterioration. Initial symptoms depend on the affected region and include dementia and movement disorders. The most common neurodegenerative brain disorders are Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, frontotemporal dementia, spinocerebellar ataxia, and Huntington's disease.



Neurodegenerative brain diseases have a strong genetic component, and Mendelian disease segregation has been documented in early-onset families. Nevertheless, the majority of the patients represent sporadic occurrences of neurodegenerative brain disease, without a familial disease history. Moreover, many instances of early-onset neurodegenerative brain disease remain genetically unexplained, making clinical diagnosis and genetic counseling difficult. A lack of knowledge about the underlying genetics impairs our ability to develop a greater understanding of the underlying pathological mechanisms, thus hindering progress in developing treatments.



SVs have already been identified in neurodegenerative brain diseases subtypes, including repeat expansions.^[3] TGS technology may therefore help investigators unravel the missing genetics in other disease subtypes as well.^[4] Long reads are advantageous for SV detection due to higher mappability in repetitive regions and their ability to span SVs entirely or to have a split alignment with a sufficiently long anchor to each side. These advantages lead to direct SV inference of breakpoints, often with nucleotide precision.^[5-7] Long-read sequencing can detect three to four times more SVs than can short-read sequencing, especially in the range of 50 to 1000 bp.

2.2 Detecting Genetic Diseases Caused by Genetic Variations

WGS technologies, including TGS, can be used to detect genetic variations in the human genome, helping to determine the causes of hereditary diseases and evaluate patient disease risks. In 2018, Miao et al.^[8] presented one of the first examples of long-read sequencing being used to identify complex structural variations in exome-negative patients, leading to successful personalized preimplantation genetic diagnosis (PGD). The study involved a patient with hepatosplenomegaly and growth retardation. Long-read sequencing detected complex intragenic repeat/deletion variations, expanding the scope of genetic testing for hereditary diseases. Long-read WGS was also performed on the patient's parents using nanopore sequencing technology at a depth of 12X. Thus, long-read sequencing provides a method for discovering genetic variations that are overlooked by short-read sequencing, which can lead to underdiagnosis or misdiagnosis of patients. Long-read sequencing thus has the potential to improve diagnostic rates in clinical settings, especially when only one pathogenic mutation is found in affected individuals suspected of carrying a recessive disease.



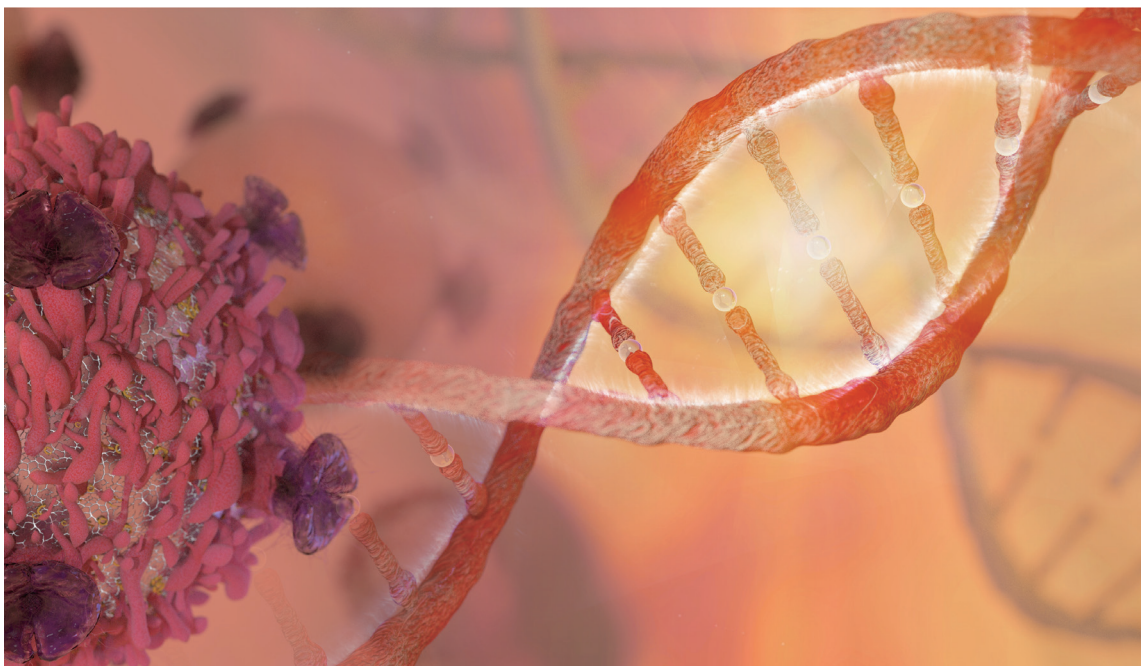
In 2022, Noyes et al.^[9] sought to better understand the whole-genome patterns of *de novo* mutations (DNMs) using long-read sequencing. They generated long-read sequence data from a family of four in which the female sibling was affected with autism no pathogenic variants were detected in NGS data alone. Deep sequencing of all four individuals was performed using three sequencing platforms (Illumina, ONT, and PacBio) and three complementary technologies (Strand-seq, optical genome mapping, and 10x Genomics). Noyes et al. initially discovered and validated 171 DNMs in two children, representing a 20% increase in the number of single-nucleotide variants (SNVs) and Indels compared to the short-read calling set. Therefore, compared to previous studies, long-read sequencing and assembly, particularly when combined with a more complete reference genome, increased the number of DNMs by over 25% and provided a more comprehensive DNM catalog compared to short-read data alone.

2.3 Detecting Cancer

WGS technology provides comprehensive information for predicting cancer drug sensitivity and prognosis, enhancing cancer prediction, diagnosis, and treatment. It addresses the challenge of identifying molecular markers in tumors with complex genomic rearrangements that lack clinical relevance. It also aids in understanding the mechanisms of resistance to chemotherapy and radiotherapy, providing important information for precision treatment.

In a 2022 study, Chae et al. performed whole genome sequencing at a depth of 10-20X on two HPV-positive and HPV-negative human samples using the PacBio long-read sequencing platform.^[10] The study identified complex genomic rearrangements resulting from HPV-mediated genomic instability in the HPV-positive case and enhancer hijacking in regions of chromosomal breaks in the HPV-negative case. These structural changes led to overexpression of cancer genes *CCND1* and *ALK*. Long-read whole genome data were also used to distinguish somatic mutations before and after structural variations, thus contributing to a better understanding of cancer evolution.

In a 2022 study, Hu et al. used SMRT and nanopore platforms to perform long-read genome and transcriptome sequencing on tumor, tumor-adjacent, and blood samples from 19 breast cancer patients.^[11] The study analyzed SVs in 28 breast cancer-related genes. The results showed that some somatic SVs appeared repeatedly in the selected genes, mostly in non-exonic regions. Hu et al. found evidence supporting the existence of SV hotspots and indicating that SV hotspots were exclusive to SNVs. Thus, the study identified SVs in breast cancer patients using long-read genome and transcriptome sequencing and demonstrated the great potential of this method in clinical applications.

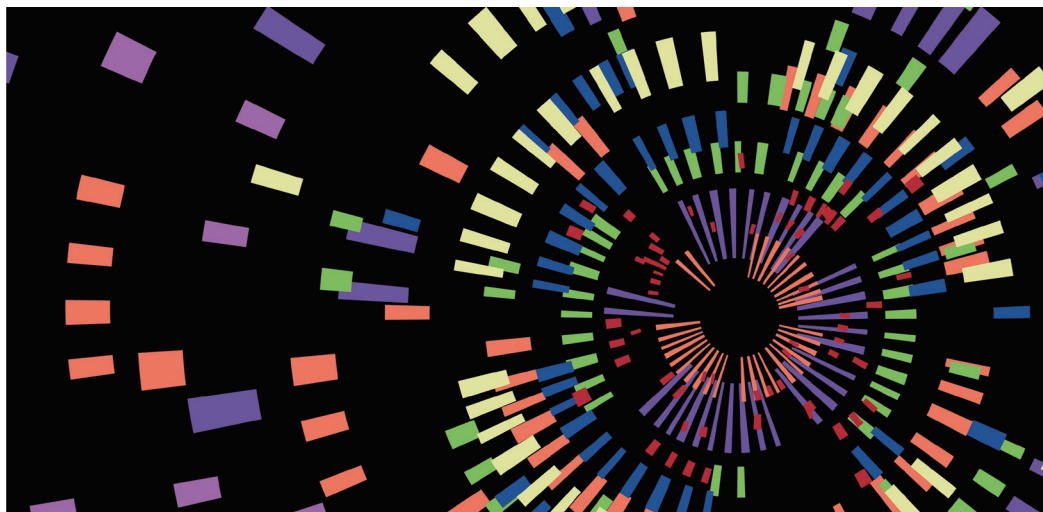


2.4 Identifying the Methylation Status of Disease-related Genes

DNA methylation-mediated gene expression regulation plays various roles in many aspects of biological processes, such as normal development and disease progression. A well-studied case involves the imprinting of the maternal or paternal genes in normal cells, where the gene locus, an array of gene loci, or even the whole X-chromosome is collectively methylated and thus silenced. In cancers, genomic DNA (gDNA) is thought to be generally hypomethylated, whereas some specific regions harboring tumor suppressor gene loci are often hypermethylated. For example, *CMYA5* was predicted to be an oncogene and a potential prognosis indicator of overall survival rate in breast cancer and might be repressed in the tumor tissue by enhanced methylation. *TSLP* is known to promote tumor cell survival and is important for metastasis in breast cancer. *TSLP* showed a lower methylation rate in the tumor tissue than that in the normal counterpart. These findings support the position that long-read methylation analysis of specimens of early-stage cancer tissues and biopsies for various cancer types would deepen our understanding of epigenomic regulation and its disturbance in cancers.^[12]

2.5 Applying Third-generation Whole Genome Low-depth Sequencing to Disease Detection

The All of Us (AoU) research team, a large genomic project of the National Institutes of Health (NIH) in the United States, recently published an article titled *Utility of long-read sequencing for All of Us* as a bioRxiv preprint.^[13] The study compared three types of sequence read data (PacBio HiFi, 6-8X; Illumina, 30X; ONT 29X) using two control samples from the AoU project. They evaluated the performance of these data on a disease-related gene set consisting of 4,641 easily sequenced genes (4,641 gene set) and 386 genes that are challenging to sequence. The study found that HiFi reads were able to accurately detect the majority of genetic variations, despite having the lowest coverage (6X-8X). In fact, among the top 10 genes ranked by F-score detected by Illumina reads (30X), HiFi reads (8X) typically achieved the same or even higher F-score values. Using HiFi reads for variant detection showed superior precision and recall rates compared to other technologies. Therefore, in certain disease detection studies, employing HiFi long-read sequencing at what is typically considered a lower depth (~10X) can be a viable strategy.



3. Conclusion

3.1 Advantages and Prospects of Long-read Whole Genome Sequencing Technology

WGS provides a wealth of genetic information and holds significant importance in the fields of human disease and oncology. In particular, WGS strategies can harness TGS-based methods to generate long sequence reads that offer promising detection approaches for discovering a broader range of SVs and variations in high-repeat regions. Combining long-read sequencing with other techniques offers significant value in diagnosing and treating diseases. Additionally, WGS has great value in pathology and therapeutics.

In the context of human disease, WGS can be used for diagnosing rare and/or genetic diseases. Analyzing WGS data allows researchers and clinicians to identify pathogenic gene mutations, determine the genetic inheritance pattern of a disease, and provide references for treatment plans. In the field of oncology, WGS can contribute to tumor genomics research by identifying genetic variation characteristics of tumors and elucidating the molecular mechanisms of different subtypes, providing a basis for tumor treatment and prognosis assessment. Genetic variations such as mutations, CNVs, and chromosomal rearrangements can be identified in tumor cells by comparing WGS data from tumor tissue and normal tissue, thus offering fundamental data for personalized treatments. Additionally, WGS can be used for population genetics and population genomics research to understand genetic variations among different populations and ethnic groups, aiding in the identification of connections between genotypes and phenotypes and providing insights into the mechanisms of human diseases.

Whether you need sample preparation, sequencing, or bioinformatics analysis, Novogene offers high-quality TGS-based WGS services such as **PacBio's Single Molecule Real-Time (SMRT) sequencing** and **Oxford Nanopore Technologies' nanopore sequencing** approaches to suit the needs of many types of experiments. As technology continues to advance, utilizing the long-read sequencing is set to play an increasingly pivotal role in conjunction with other technologies such as single-cell analysis and epigenome results. Building on years of sequencing expertise and efficient standard workflow, Novogene offers versatile solutions for multiple biological inquiries, spanning genomics, transcriptomics, epigenomics, and metagenomics.



References:

1. Adewale BA Will. long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *Afr J Lab Med*. 2020;9(1), a1340.
2. Logsdon GA, Vollger MR, and Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020 Oct;21(10):597-614.
3. Pihlstrøm L, Wiethoff S, and Houlden H. Genetics of neurodegenerative diseases: an overview. *Handb Clin Neurol*. 2017;145:309-323.
4. De Coster W, and Van Broeckhoven C. (2019). Newest Methods for Detecting Structural Variations. *Trends Biotechnol*. 2019 Sep;37(9):973-982. Epub 2019 Mar 19.
5. Treangen TJ and Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet*. 2011 Nov 29;13(1):36-46.
6. Sedlazeck FJ, Lee H, Darby CA, and Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet*. 2018 Jun; 19(6): 329-346.
7. Lee H, Gurtowski J, Yoo S, Nattestad M et al. Third-generation sequencing and the future of genomics. 2016 *bioRxiv*. doi.org/10.1101/048603.
8. Miao H, Zhou J, Yang Q, Liang F et al. Long-read sequencing identified a causal structural variant in an exome-negative case and enabled preimplantation genetic diagnosis. *Hereditas*. 2018 Sep 28;155:32.
9. Noyes MD, Harvey WT, Porubsky D, Sulovari A et al. Familial long-read sequencing increases yield of de novo mutations. *Am J Hum Genet*. 2022 Apr 7;109(4):631-646.
10. Chae J, Lee JS, Park J, Lee D-S et al. Deciphering the Evolutionary History of Complex Rearrangements in Head and Neck Cancer Patients Using Multi-Omic Approach. 2022 *bioRxiv*. doi.org/10.1101/2022.08.19.504509.
11. Hu T, Li J, Long M, Wu J et al. Detection of Structural Variations and Fusion Genes in Breast Cancer Samples Using Third-Generation Sequencing. *Front Cell Dev Biol*. 2022 Apr 13;10:854640.
12. Sakamoto Y, Zaha S, Nagasawa S, Miyake S et al. Long-read whole-genome methylation patterning using enzymatic base conversion and nanopore sequencing. *Nucleic Acids Res*. 2021 Aug 20;49(14):e81.
13. Mahmoud M, Huang Y, Garimella K, Audano PA, et al. Utility of long-read sequencing for All of Us. 2023 *bioRxiv*. doi.org/10.1101/2023.01.23.525236.


Novogene is committed to be **Your Trusted Genomics Partner**

- Trusted International Corporation
- Trusted Comprehensive Experience
- Trusted Expertise & Technology
- Trusted Service & Standardized Procedures
- Trusted Delivery Quality & Efficiency



Follow us on LinkedIn


Novogene Co., Ltd

 www.novogene.com

 [Novogene Global](#)

 [Novogene Global](#)

 [Novogene Global](#)

 Inquiry_us@novogene.com | info@novogene-europe.com | marketing_amea@novogeneait.sg

© 2011-2023 Novogene Co., Ltd. All Rights Reserved

Information and specifications are subject to change at any time without notice. Please contact your Novogene representative.