

UTRECHT UNIVERSITY

Writing Assignment Thesis

---

---

Using Long-Read Sequencing Techniques to Overcome Limitations of  
Routine Genetic Testing: Unraveling the Complexity of the SMN Locus

---

---

*Author:*

Onur Mert Batmaz

*Daily supervisor:*

Maria Zwartkruis

*Examiner:*

Gijs van Haften

*Second reviewer:*

Hanneke van Deutekom

Van Haften Group

Department of Genetics, University Medical Center Utrecht

## Layman's Summary

The blueprint of life is hidden in a double-stranded, helical structured molecule called DNA and is written with an alphabet that consists of four types of nucleotides. The order of these nucleotides ultimately determines the structure of proteins and causes cells to function properly. DNA sequences can undergo alterations as small as a single nucleotide to as big as hundreds of thousands of nucleotides. These alterations could lead to changes in protein structure and development of diseases. Sequencing and identifying genomic alterations can be used in clinical studies to understand disease development as well as for treatment. The first standard in genome sequencing, Sanger method, was expensive and slow. Limitations of Sanger sequencing have led to the development of Next Generation Sequencing (NGS) technologies. NGS methods break DNA molecules into small fragments and use the overlapping regions between the fragments to provide fast, cheap, and accurate genome sequences for analysis. NGS methods are the current golden standard used in clinical studies. Although, NGS techniques are suitable for detecting small changes, they cannot detect large alterations or regions with repetitive sequences, which are referred to as camouflaged. Because of NGS methods' inability to analyze camouflaged regions, long-read sequencing (LRS) methods have been established. LRS technology is shown to be capable of sequencing repeated regions and large structural alterations. Therefore, LRS methods are used in clinical studies to analyze DNA sequences with more detail that cannot be achieved with NGS techniques. Spinal muscular atrophy (SMA) is a neuromuscular disease that causes progressive loss of muscle control, movement, and strength due to defects in *survival motor neuron (SMN)* gene. *SMN* locus contains two almost identical genes, *SMN1* and *SMN2*, and the loss of *SMN1* is shown to cause SMA. *SMN* genes can undergo changes that lead to mixed (hybrid) *SMN1-SMN2* sequences. Individuals can also have different copy numbers of *SMN2*. The relationship between genetic sequence variation and the severity of the SMA phenotype also has not been clearly established, although *SMN2* copy number is used to make rough predictions. Such reasons contribute to the complexity of the *SMN* locus, and for decades, clinical studies have tried to sequence *SMN* locus for its relevance to SMA. However, NGS methods failed to resolve this region because of its structural complexity. Since LRS methods can overcome the limitations of NGS, they can be used in the analysis of *SMN* locus to identify organizational and structural variations that couldn't be determined by NGS methods. LRS methods can also possibly improve SMA diagnosis by identifying the pattern of genomic variations that are associated with the severity of SMA phenotypes. Furthermore, LRS can be potentially used to improve genetic screening for the detection of SMA carriers. Consequently, LRS can be clinically used not only in diagnosis of SMA patients, but it can also be used in genetic counseling of individuals with a high risk of having a child with SMA.

## Acknowledgements

I would like to thank Maria Zwartkruis for her assistance and supervision. I wish to extend my special thanks to Gijs van Haaften and Hanneke Deutekom for giving me the chance to write this literature review.

## List of Abbreviations

FGS = first-generation sequencing  
Bp = base pair  
NGS = Next-generation sequencing  
TRs = Tandem repeats  
SVs = Structural variations  
SNV = single nucleotide variations  
LRS = long-read sequencing  
SMRT = single-molecule real-time sequencing  
Pacific Biosciences = PacBio  
Kilobases = kb  
Megabases = Mb  
Oxford nanopore sequencing= ONT  
SMN = Survival motor neuron  
SMA = Spinal muscular atrophy  
FL-SMN = full-length SMN  
VNTR = variable number of tandem repeats  
qPCR = quantitative polymerase chain reactions  
MLPA = multiplex ligation-dependent probe amplification  
aCGH = microarray-based comparative genomic hybridization  
HRG = Human reference genome  
PBS = Poretti-Boltshauser syndrome  
Plastin 3 = PLS3

# TABLE OF CONTENTS

---

- Abstract ..... 1
- Introduction..... 1
  - Sequencing Strategies and Methods..... 4
    - Basic Principals of Sequencing Approaches ..... 4
    - Short-read and Long-read Sequencing Methods ..... 5
    - NGS Methods..... 6
    - Advantages and Limitations of NGS Methods..... 7
    - LRS Methods..... 8
  - Applications and Advantages of LRS in Clinical Studies ..... 11
  - Resolving Complex SMN Locus by LRS Methods ..... 14
- Discussion ..... 16
- Conclusion ..... 19
- References..... 19
- Appendix A ..... 26

# Using long-read sequencing techniques to overcome limitations of routine genetic testing: Unraveling the complexity of the SMN locus

## ABSTRACT

---

The field of genetic sequencing is rapidly expanding, with significant consequences for research and clinical practice. Next-generation sequencing (NGS) technology allows for accurate, fast, and cost-effective genome sequencing. NGS is suitable to detect small-scale DNA alterations but not for substantial structural variations (SVs) and tandem repeats (TRs). As a result, NGS methods rely on additional tests to analyze such complex regions. Long-read sequencing (LRS) technology is developed in response to the limitations of NGS technologies and is capable of sequencing complex genomic regions without requiring additional tests. Spinal muscular atrophy (SMA) is a neuromuscular disease caused by defects in the *survival motor neuron (SMN)* locus. NGS has been used to analyze the *SMN* locus for decades. However, the relationship between genetic alterations within *SMN* locus and SMA phenotype variety is still not completely understood. LRS has been used to resolve complex regions and has identified a wide range of clinically relevant structural variants. This literature review shows that LRS could also be similarly used to analyze *SMN* locus. Furthermore, LRS could potentially identify genomic alterations that contribute to SMA phenotype variability and could be used as a genetic screening method to detect SMA silent carriers. Overall, the use of LRS in clinical studies could improve SMA diagnosis as well as be used in genetic counseling of individuals with a high risk of having a child with SMA.

## INTRODUCTION

---

Accurately sequencing DNA has been the ultimate goal for many since the realization of DNA holds the blueprint of life. Sanger sequencing method successfully sequenced DNA for the first time<sup>1</sup>. Later on, this groundbreaking technology has inspired the concept of first-generation sequencing (FGS) platforms. In the following two decades, advances in sample preparation, signal detection, and base calling have made Sanger sequencing the traditional method for genome sequencing. These improvements, combined with the automation of the FGS workflows, enabled the initiation of the Human Genome Project<sup>2</sup>.

However, at the end of this project, it was evident that FGS methods had reached their technical threshold. These methods were not cost-effective and too slow for clinical diagnosis<sup>2,3</sup>. Furthermore, the accuracy of the FGS reads significantly dropped after 900 base pairs (bp)<sup>4</sup>. Because of these limitations, Next-Generation Sequencing (NGS) techniques have emerged as an improved alternative to FGS for genome sequencing.

NGS is an umbrella term for short-read sequencing platforms. The three leading NGS technologies are Illumina sequencing by Illumina Inc., Ion Torrent sequencing by ThermoFischer Scientific, and Nanoball sequencing by Beijing Genomics Institute. Each platform has released various instruments that differ by throughput and read length. However, all NGS platforms share three key characteristics in their library preparation process<sup>2</sup>. One, fragmentation of the target DNA/RNA molecules. Two, ligation of short adaptor sequences on the fragmented molecules to induce solid surface attachment or circularization. Three, the amplification of the fragments.

NGS platforms are proven to be very useful in both fundamental and clinical research because of their high throughput, high accuracy, cost-effectiveness, and rapid sequencing ability<sup>2,3,5-7</sup>. Nonetheless, these platforms often produce ambiguous reads from genomic regions that contain high GC content, tandem repeats (TRs), and structural variations (SVs, >50 bp)<sup>2,3,5-7</sup>. Therefore, studies that use NGS techniques often overlook such genomic regions. Furthermore, NGS short-read libraries mix parental sequences, which makes NGS methods unsuitable for haplotype phasing<sup>7</sup>. The limitations of NGS platforms have driven the development of third-generation long-read sequencing (LRS) platforms.

Single-molecule real-time sequencing (SMRT) by Pacific Biosciences (PacBio) is one of the leading LRS technologies in genomic studies. PacBio sequencing produces reads that are approximately 20 kilobases (kb) long<sup>2,3,6,7</sup>. Nanopore sequencing by Oxford Nanopore Technologies (ONT) is another LRS technology used in genome sequencing. ONT platforms are capable of producing ultralong DNA reads. These reads range from hundreds of kb to several megabases (Mb)<sup>2,3,6-8</sup>. However, both LRS platforms suffer from high inaccuracy (~%14)<sup>2,3,6,7</sup>. Furthermore, Bionano Genomics has established a novel high-throughput, native, single-molecule level mapping technology that allows *de novo* assembly of genomes with very high accuracy<sup>2,3,7,9</sup>. By covering large segments of the DNA, LRS offers a novel approach to resolving complex regions of the genome.

An example of such a complex region is the *survival motor neuron (SMN)* locus. Spinal muscular atrophy (SMA) is one of the most commonly occurring autosomal-recessive diseases with an incidence rate of 1 in 10,000 births and is associated with the *SMN* locus<sup>10,11</sup>. This locus contains two homozygously

expressed genes: *SMN1* and *SMN2*. Approximately 95% of SMA patients have lost their *SMN1* gene either through homozygous deletion or by a gene conversion event<sup>11–19</sup>. The remaining 5% of SMA patients, on the other hand, contain intragenic mutations in their *SMN1* gene that cause aberrant SMN protein production<sup>11,19–21</sup>.

Surprisingly, *SMN2* is almost an exact copy of *SMN1*, except for 16 positions involving a total of 20 nucleotides<sup>22</sup>. The variant, c.840C>T located in exon 7, especially plays a crucial role in the *SMN2* expression<sup>23</sup>. Because of this variant, most *SMN2* transcripts exclude exon 7; whereas, only a fraction of the *SMN2* transcripts retain their exon 7 to produce full-length SMN (FL-SMN) protein. Another variant that occurs in intron splicing site 1 also causes the exclusion of exon 7, further contributing to the deficiency of FL-SMN<sup>24</sup>. Therefore, SMA patients with *SMN1* deletion or loss-of-function mutations can only express residual levels of FL-SMN through their *SMN2* genes. The insufficient expression of FL-SMN causes degeneration and loss of anterior horn cells in the spinal cord and brainstem nuclei<sup>25</sup>. These pathological changes ultimately lead to muscle weakness and atrophy.

The exact genomic mechanisms dictating the severity of SMA symptoms are still unknown. Since each copy of *SMN2* produces a residual amount of SMN protein, *SMN2* copy number variations are used in *SMN* genetic screening to assess the expected symptomatic severity of SMA patients to a certain degree. SMA patients with low *SMN2* copies often exhibit severe SMN protein deficiency and are associated with severe SMA symptoms. In contrast, due to their relatively elevated SMN protein levels, SMA patients with high *SMN2* copies are often associated with milder SMA symptoms<sup>11,26,27</sup>. Nonetheless, a large fraction of SMA patients exhibits pathogenic symptoms that are not correlated with their *SMN2* copy number<sup>27</sup>. These observations indicate that other currently unknown genetic variations besides *SMN2* copy number determine the symptomatic outcome of SMA.

Resolving the *SMN* locus is clinically relevant because of the reasons mentioned above. Yet, the current NGS techniques fail to unravel the complexity of the *SMN* locus due to their innate technical limitations. For instance, using short-reads to map *SMN* locus yields ambiguous regions due to the sequence homology between *SMN1* and *SMN2*. Moreover, the *SMN1-SMN2* gene conversion events lead to the formation of hybrid genes, which further contributes to the complexity of the *SMN* locus<sup>19,28</sup>. NGS mapping algorithms often fail to detect the extent of sequences involved in such gene hybridization events. In addition, the *SMN2* gene is an inversion of *SMN1*<sup>29</sup> and short-read libraries are not suitable for establishing the exact breakpoints for such copy-neutral variations<sup>30</sup>. Lastly, NGS platforms cannot distinguish the silent *SMN1* mutation carriers (2 + 0) from the healthy individuals and can yield false-

negative results<sup>28,31,32</sup>. LRS techniques, on the other hand, are capable of resolving the *SMN* locus as they do not share the same technical limitations with NGS techniques.

This paper has two aims. First, to show long-read sequencing platforms offer an alternative in genome sequencing that can overcome the limitations of NGS platforms. Second, to explain that the routine genetic testing methods cannot resolve the *SMN* locus. The LRS and mapping methods such as PacBio HiFi, ONT, and Bionano are more suitable to unravel this complex locus.

## SEQUENCING STRATEGIES AND METHODS

The NGS and third-generation long-read sequencing (LRS) platforms offer vast possibilities of sequencing approaches and capabilities. Selecting sequencing strategies and platforms accordingly to the goals of the study optimizes the quality and accuracy of the results. Therefore, what type of sequencing strategy and platform to use should be carefully considered before the experiment. Targeted sequencing, whole-exome sequencing (WES), and whole-genome sequencing are the three sequencing methodologies based on the size of the genetic area covered (WGS). Whereas, based on the length of reads produced, sequencing platforms are classified as NGS or LRS.

### Basic Principals of Sequencing Approaches

The sequencing strategy involves deciding how much of the genome to examine. Depending on the study's goals, the region to be sequenced can be as small as a few exons or as large as the entire genome. Targeted sequencing approach detects and sequences genomic regions that contain pre-determined specific sequences called motifs<sup>2</sup>. WES and WGS, on the other hand, are more inclusive sequencing techniques. WES method recognizes and sequences all protein-coding regions (exons) in the genome<sup>2</sup>. In addition to exons, WGS sequence the entire genome, including the non-coding regions (introns) as well<sup>2</sup>.

Targeted sequencing strategy is one of the most commonly used methods in clinical gene panels as they offer a cheap and fast genetic analysis option<sup>2</sup>. Because it only covers a small part of the genome, targeted sequencing also requires less computational work. However, one of the disadvantages of this approach is its requirement of prior knowledge of the sequence of interest. Therefore, targeted

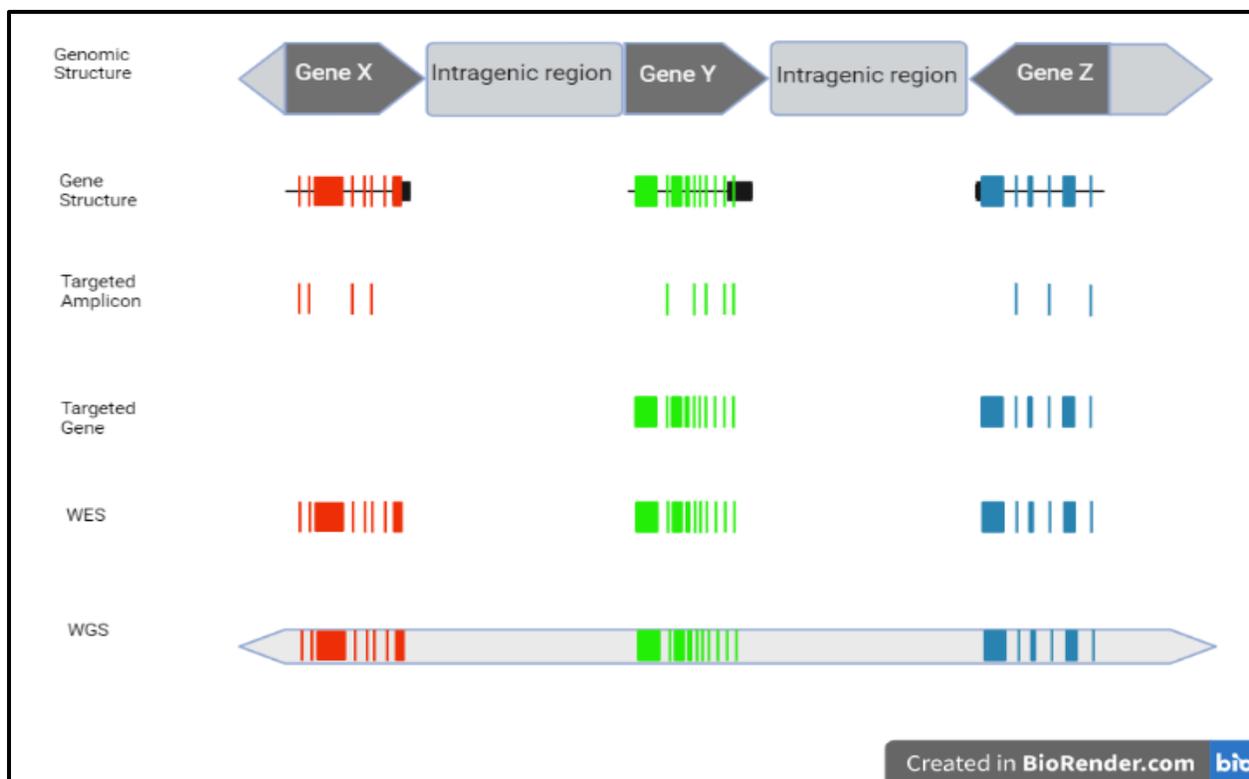
sequencing is not suitable for studies involving the analysis of 'camouflaged' genomic regions<sup>2</sup>. Furthermore, targeted sequencing has the risk of underrepresenting genetic information due to strictly relying on motifs. For instance, if the used motif is not present in every exon of a gene, those exons without the target motif will not be recognized as part of the gene.

WES is another method used for genomic screening in clinical studies. This approach is especially beneficial considering exons contain approximately 85% of all known disease-associated variants<sup>2,33,34</sup>. Principles of using WES comes from important role of proteins in cellular processes. Proteins are synthesized through the exon sequences and any change in these exons can lead to aberrant protein formation. The function of a protein comes from its structure, because of this, aberrant proteins cannot function properly. As a result, cellular process cannot be completed with aberrant proteins, which can lead to pathogenesis. However, exons are estimated to make up 2% of the whole human genome<sup>2,35</sup>. Therefore, the disadvantage of WES is the bulk of the human genome cannot be reached for analysis by this method.

Although exons are very important for their role in protein synthesis, introns should not be overlooked. Besides their role in genome stabilization, introns also contain gene regulatory sites, ribosomal RNA sequences, and disease-associated variants<sup>36,37</sup>. Therefore, sequencing introns is clinically relevant as much as sequencing exons. WGS offers a thorough clinical analysis option by sequencing both introns and exons. Regardless, this also makes WGS the most expensive, time-consuming, and computationally work-intensive approach. The figure below illustrates each gene sequencing approach (fig. 1).

### Short-read and Long-read Sequencing Methods

As previously stated, the chosen sequencing platform must be capable of following the study's objectives. This is due to the fact that each platform has different technical limitations caused by variations in sample preparation, signal detection, and sequence analysis methods. If the technical applicability of the sequencing platform is not suitable for research purposes, it will adversely affect the genomic analysis efforts. Therefore, the advantages and disadvantages of sequencing platforms should be considered carefully. Specifications of different sequencers can be found in table 1 appendix A.



**Figure 1 - Genome sequencing strategies.** The top line illustrates a gene cluster with 3 genes (gene X, Y and Z). The second line represents the structure of each gene, where exons are represented by colored areas and introns are indicated by black lines. Targeted amplicon sequencing is only able to sequence exons with pre-determined motif in it. Targeted gene sequencing approach only sequence the pre-determined genes. WES approach sequences all of the exons; whereas, WGS approach sequence both exons and introns. (Created with BioRender.com)

## NGS Methods

Illumina sequencing technology is the market-leading platform in the field of NGS. During Illumina library preparation, DNA molecules are fragmented into ~500 bp long pieces, ligated with adapter sequences, and bound to glass slides. Then, glass-bound fragments are PCR amplified to form clusters that contain millions to billions of clonal DNA templates<sup>2,6</sup>. Illumina sequencing technology utilizes fluorescently labeled reversible terminator nucleotides during its sequencing-by-synthesis step. Depending on their identity, nucleotides release distinctively colored lights as signals when they are incorporated; while, their modified terminator groups prevent the incorporation of more nucleotides. Once the high-resolution optical imaging device records the signal, the terminator group of nucleotides is released, and the cycle ends. This process is repeated until a read with a length of 75-300 bp is generated<sup>2,6</sup>.

Another important NGS platform is BGI's Nanoball Sequencing technology. In this technique, DNA molecules are fragmented into ~300 bp long pieces and ligated with adapter sequences. Then, split oligonucleotides are attached, and the fragments are denatured. Single-stranded DNA molecules are circularized, chain amplified by Phi 29 DNA polymerases to form nanoballs and anchored to patterned flow cells<sup>2</sup>. Similar to Illumina Sequencing, Nanoball Sequencing technology also uses the sequence-by-synthesis approach. Nucleotides modified with distinct fluorophore groups are released into flow cells to be incorporated. The remaining nucleotides are washed off from the flow cells after one of them is incorporated. A laser excites the newly incorporated nucleotide to emit fluorescent light as a color. High resolution charge-coupled device (CCD) cameras then detect these as distinctive base calls depending on the color of the light. This nucleotide incorporation cycle repeats another ~150 times<sup>2</sup>.

Another significant NGS technique is ThermoFischer Scientific's Ion Torrent sequencing. DNA molecules are initially fragmented into 200 bp to 1500 bp long sections in this procedure. Short adapter sequences are ligated to the fragments to anchor them to specially designed beads. Bead-bound fragments are amplified via emulsion PCR and then placed into an array of wells that contain highly sensitive semiconductor chips<sup>2</sup>. Unlike the other NGS techniques, Ion Torrent Sequencing does not use fluorescently labeled nucleotides. Instead, it utilizes semiconductor chips to detect label-free base incorporation. At each cycle, wells receive one type of nucleotide at a time until base incorporation occurs. When the incorporation of a base occurs, the reaction releases a hydrogen atom from the polynucleotide backbone. The released Hydrogen ions generate small currents within the wells. The highly sensitive semiconductor chips then detect and record these currents as base calls. This cycle repeats reads are 200 bp - 600 bp long<sup>6</sup>.

### Advantages and Limitations of NGS Methods

All NGS platforms utilize parallel sequencing, albeit through different approaches. Because of this, NGS platforms have very high throughput and are ideal for fields that require fast genome sequencing<sup>2</sup>. Furthermore, NGS platforms generate large libraries that consist of thousands to millions of short-read sequences. Specially designed algorithms use regions over-lapping between the short sequences to map a final consensus sequence from these large libraries. Through these consensus sequences, NGS platforms distinguish small variants from sequencing artifacts with high confidence. As a result, NGS platforms are advantageous for detecting single nucleotide variations (SNVs) and indels (<50 bp) with high accuracy<sup>2</sup>.

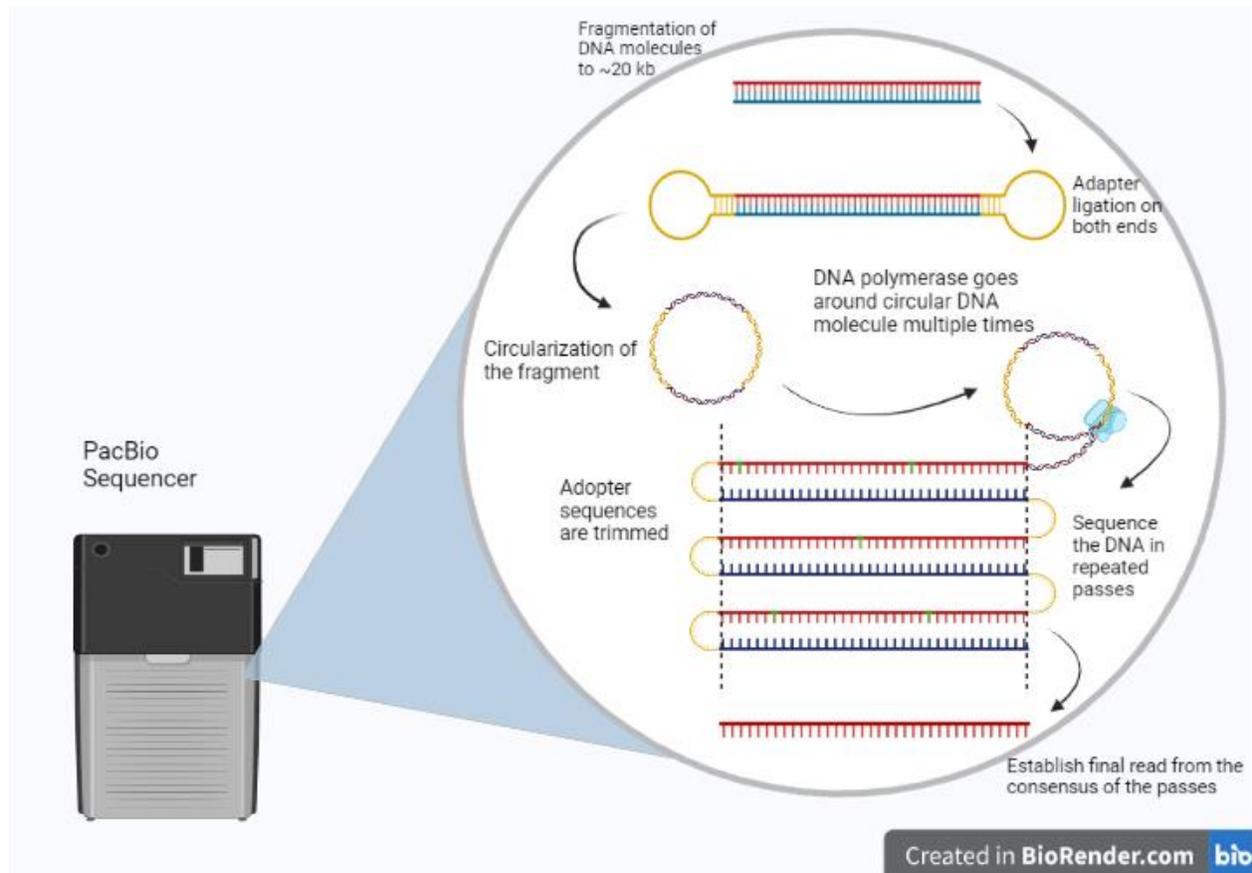
Another advantage of NGS platforms is that they are very cost-effective. Ion Torrent Sequencing is especially a cheaper option among the NGS technologies since it does not rely on sophisticated light sources, scanners, or cameras. Another advantage of NGS platforms is that they are very cost-effective<sup>2</sup>. Ion Torrent Sequencing is especially a cheaper option among the NGS technologies since it does not rely on sophisticated light sources, scanners, or cameras. Moreover, many bioinformatics tools and algorithms, such as GATK and DeepVariant, are designed to be used by NGS platforms<sup>2</sup>. The availability of a broad spectrum of tools and algorithms allows users to analyze their data in the most optimum way possible.

Even though NGS techniques are advantageous in many ways, they also have innate technical limitations. One disadvantage of NGS platforms is their inability to resolve large TRs and SVs (>50 bp)<sup>38</sup>. These variants include variable number of tandem repeats (VNTR), homopolymer sequences, repeat expansions, additions, deletions, inversions, and translocations<sup>2,5,6</sup>. These types of long TRs and SVs are problematic for NGS platforms because they exceed the boundaries of the reads used in their libraries. For this reason, NGS algorithms fail to establish the endpoints of these variants and underrepresent them, causing a sequencing bias. Furthermore, NGS libraries contain reads from both maternal and paternal chromosomes. NGS methods are not suitable for studies that aim to do haplotype phasing since they cannot distinguish the parental fragments from each other. NGS platforms are also disadvantageous for sequencing regions with high GC- content due to their use of PCR during library preparation<sup>38</sup>.

## LRS Methods

PacBio SMRT sequencing is the first popularly used LRS method in genomic studies<sup>39</sup>. The initial PacBio SMRT sequencing workflows have used 1kb long, linear DNA molecules. However, the improvements in PacBio sample preparation have increased this length to ~20 kb<sup>2,3</sup>. PacBio sequencing shares similarities to NGS techniques in the use of adapter sequences and fluorescently labeled nucleotides. However, PacBio uses hairpin adapters to connect and circularize the DNA molecules. Zero mode waveguides disperse these circular DNA molecules into nanowells containing fluorescently labeled nucleotides and a DNA polymerase<sup>2,3,5</sup>. Then, polymerases anchor the template DNA molecules to the bottom of the wells and initiate synthesis. The CCD cameras at the bottom of the wells record the colors emitted from the incorporated nucleotide as base calls. Because of their circular structure, polymerases can sequence template DNA molecules several times. The overall PacBio High Fidelity (HiFi) sequencing

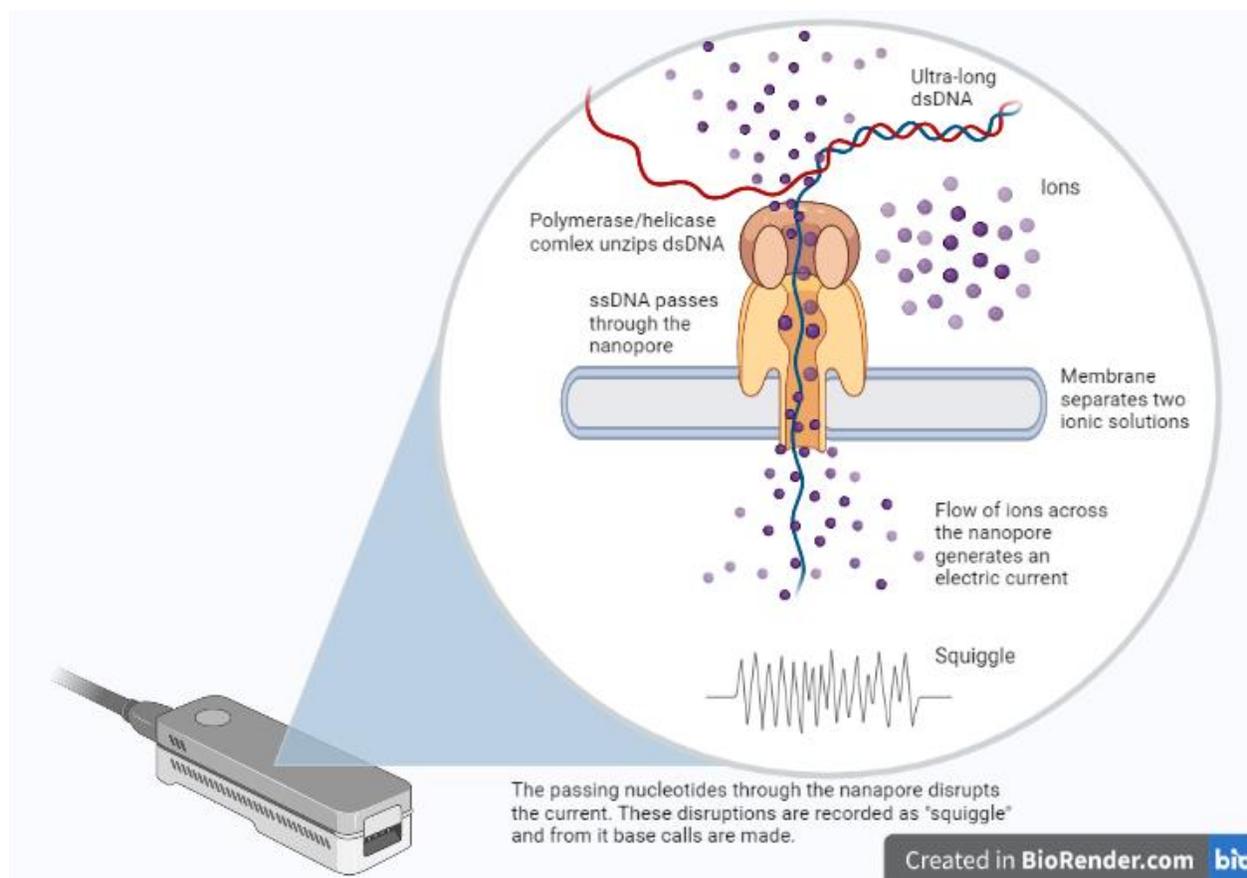
process is illustrated in figure 2. PacBio SMRTbell sequencing has improved its accuracy from 87% to >99% by establishing such consensus sequences<sup>40,41</sup>.



**Figure 2 - PacBio HiFi Sequencing.** Fragmented, ~20 kb long dsDNA strands are ligated with adapter sequences to give them circular shapes. Through their circular shapes, DNA fragments can be sequenced in multiple passes. Algorithms detect the adapter sequences connecting the DNA sequences and trim them out from the final sequence. The wrongly sequenced bases within the passes are indicated with green color. The consensus of all passes are used to generate a final sequence, which increases the base calling accuracy >99%. (Created with BioRender.com)

The other popularly used LRS platform is ONT. One of the unique aspects of this technology is its use of a membrane with nanopores on it. This membrane generates an electric current by separating the ionic solution into two sides. A phi-29 DNA polymerase/helicase protein complex is also attached to the nanopores<sup>2,3</sup>. Another unique aspect of ONT is its ability to sequence native, ultra-long DNA molecules. Through adapter sequences, ultra-long double-stranded DNA molecules interact with nanopore complexes. The phi-29 DNA polymerase/helicase protein complex separates the DNA strands, while the motor protein feeds one of the strands through the nanopores. The movement of the nucleotides across

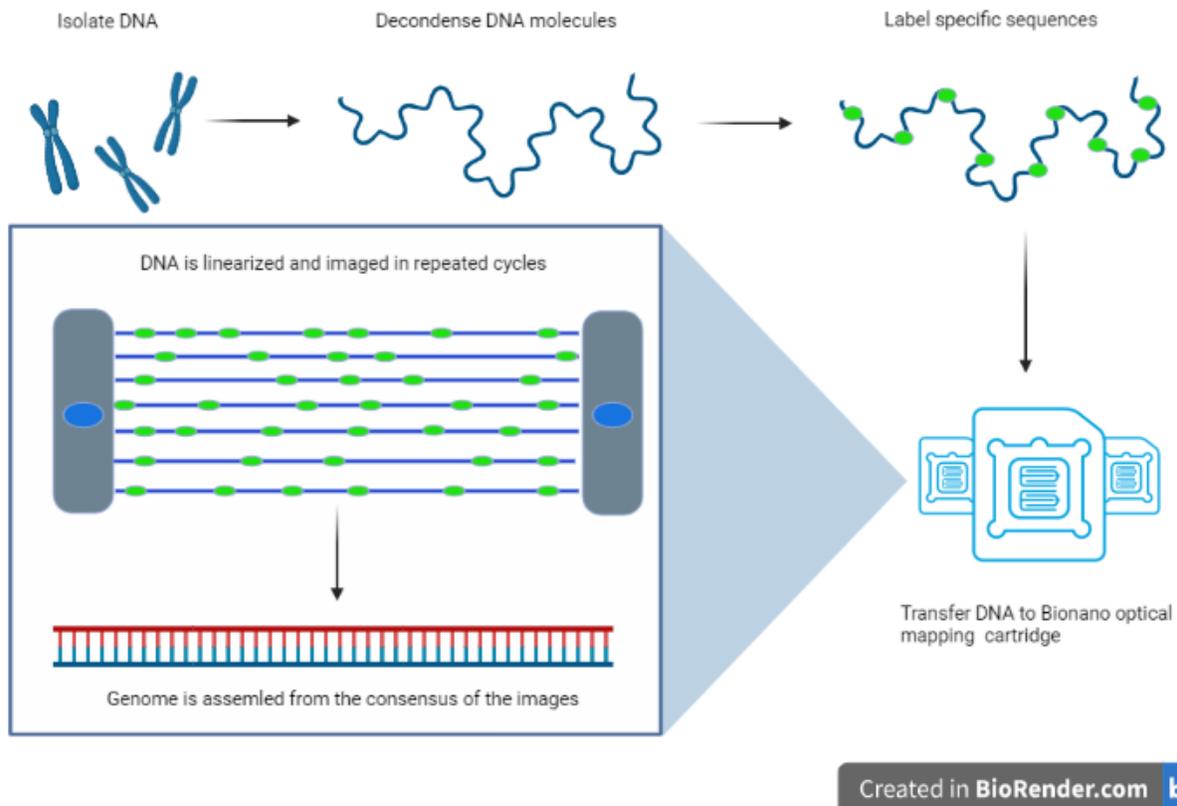
the nanopore disrupts the current, and sensitive chips register these distinctive disruption patterns as base calls<sup>2,3,42</sup>. Figure 3 briefly illustrates the process of ONT sequencing. Unlike other sequencing techniques, the limiting factor for ONT is the length of the DNA molecule. Although the current ONT read lengths are several Mb long, this platform is capable of sequencing DNA in gigabase range in theory<sup>38</sup>.



**Figure 3 - ONT Sequencing.** Ultra-long, double-stranded DNA is ligated with adapter sequences. Through these adapter sequences DNA molecules attach and separated. The template strand is fed into the nanopore by the motor protein while flow of ions across the membrane generates an electric current. The movement of the nucleotides disrupts the electric current, where these distinctive disruption patterns are recorded as base calls. (Created with BioRender.com)

Bionano also uses LRS data to offer relatively cheaper, high-throughput optical mapping technology for genome assembly<sup>43</sup>. Even though this technology does not have a single-nucleotide resolution, it can detect SVs as small as 500 bps<sup>44</sup>. Because of this, Bionano optical mapping technology can be used in *de novo* genome assembly studies as well<sup>43</sup>. This method uses DNA sequences as targets to fluorescently label ultra-long DNA fragments. Then, labeled DNA molecules are linearized in parallel and

imaged by specially designed chips. Tagged DNA regions combined by the chips to form a final consensus sequence<sup>45</sup>. The image below illustrates the optical mapping process (fig. 4). By imaging whole intact single molecules of DNA in their native states, Bionano optical mapping technology can accurately identify SVs and CNVs. Furthermore, variant annotation pipeline can screen among the detected variants to find the ones relevant to the disease or phenotype of interest<sup>43</sup>.



**Figure 4 - Bionano optical mapping.** Isolated high molecular weight DNA are labeled at specific sequences. Consecutive images of labeled DNA molecules are taken inside the specially designed chips. Algorithms combine these images to assemble the final sequence. (Created with BioRender.com, image is adopted from <https://bionanogenomics.com/research/genome-assembly/>)

## APPLICATIONS AND ADVANTAGES OF LRS IN CLINICAL STUDIES

Clinical studies screen for SVs, TRs, and copy number variations (CNVs) for four reasons. One benefit of these variations is that they shed light on the genetic causes of diseases. Two, clinical tests use these differences to predict phenotypic outcomes. Three, identifying disease-related genomic variations

allows for clinical diagnosis, prognosis, and treatment options. Four, these variations are used in genetic screening tests to identify people at risk of developing diseases.

As previously mentioned, NGS methods are not suitable for de novo genome assembly, haplotype phasing, and SV detection<sup>41</sup>. Therefore, studies that use NGS also employ other techniques to verify their results. Some of the methods include Southern blot, quantitative polymerase chain reactions (qPCR), multiplex ligation-dependent probe amplification (MLPA), and microarray-based comparative genomic hybridization (aCGH)<sup>28,30,46,47</sup>. Yet, relative to the LRS methods, these techniques still provide limited information about genomic variations<sup>30,40,46,48</sup>.

CNVs are associated with various types of diseases such as cancer and neurodegenerative diseases. Dosage-sensitive methods such as qPCR, MLPA, and TaqMan are used in CNV analysis<sup>49</sup> but fail to distinguish the CNVs between the parental chromosomes<sup>28,31,32</sup>. MLPA and qPCR techniques provide a robust and sensitive way to assess copy numbers but must be locus-specific<sup>46</sup>. Although hybridization-capture-based methods can overcome this limitation, they do not offer high sensitivity<sup>46</sup>. On the other hand, LRS overcomes the limitations of these routinely used tests by providing information about the nucleotide content, location, and length. For instance, targeted Nanopore sequencing identified CNVs of 66 genes in plasma samples taken from lung cancer patients to generate patient profiles<sup>50</sup>. This study shows that not only LRS is highly suitable for CNV analysis, but with customized workflows, it can become suitable for cell-free DNA analysis as well.

TRs are also associated with neurological diseases, rare diseases, and common complex disorders<sup>38,51,52</sup>. Southern blot and repeat-prime PCR techniques are two of the most widely used techniques for TR analysis. Although Southern blot is suitable for determining the length of short TRs, it omits information on the nucleotide composition of the repeats. The accuracy of this technique also decreases as the length of the repeats increase. Other limitations of Southern blot include high turnaround time and locus-specific analysis<sup>40,42</sup>. Repeat-prime PCR can determine the nucleotide content of the TRs but not their location within the genomes<sup>42</sup>.

Unlike NGS, Southern blot and repeat-prime PCR techniques, LRS does not share these limitations. While traditional techniques fail when used alone, LRS methods can successfully identify the nucleotide content, size, and distribution pattern of TRs associated with diseases without requiring other techniques for verification<sup>38</sup>. Moreover, because of their non-locus-specific approach, LRS techniques can identify TRs in multiple loci<sup>42</sup>. Due to the reasons mentioned above, LRS approaches are used in numerous studies to analyze the involvement of TRs in several neurodegenerative disorders. Some of these diseases include

Fragile X-associated tremor/ataxia syndrome, neuronal intranuclear inclusion disease, oculopharyngeal muscular dystrophy, spinocerebellar ataxia, and Huntington's disease<sup>38,53–57</sup>.

SVs are seen in four forms: duplications, deletions, inversions, and translocations. Some traditionally used SV analysis methods involve Southern blot, fluorescent in situ hybridization, and pulse-field gel electrophoresis. However, these techniques require prior knowledge of the sequence desired to be analyzed. Because of this, these methods are not suitable for studies aiming to identify novel SVs<sup>30</sup>. On the other hand, aCGH can identify SVs without requiring prior knowledge of their sequences. However, aCGH cannot determine the location as well as the orientation of SVs<sup>30,46,58</sup>. In contrast to aCGH, strand-seq can detect the orientation of inverted sequences but cannot identify their breakpoints<sup>58</sup>. As shown above, techniques commonly used in SV analysis with NGS are not comprehensively capable of capturing the complexity of SVs.

Conversely, LRS methods are able to resolve complex SVs, regardless of their type. For instance, targeted capture nanopore sequencing has identified 11 distinctive SVs in patients with Lynch syndrome. The disease-associated deletions and duplications are found in the mismatch repair genes *MLH1* and *MSH2*. By accurately identifying the SVs and their breakpoints, the study has shown that targeted Nanopore sequencing for multiple genes is useful for the identification of pathogenic SVs and might replace gene panels for the screening of SVs in hereditary diseases<sup>59</sup>. In a separate investigation, Bionano Genomics mapping technology has revealed a 90-kb deletion inside the *CDKL5* gene. Because of its mosaic structure, this novel SV has never been detected by chromosomal microarrays before. However, genetic panels can employ this SV as a prospective diagnostic marker for X-linked dominant early infantile epileptic encephalopathy-2 using LRS techniques<sup>60</sup>. LRS methods are also capable of detecting inversions. A pathogenic 12 kb inversion within the *BRPF1* gene is identified by PacBio Sequencing. This inversion, which disrupts the gene activity, is linked to intellectual developmental disorders with dysmorphic facies and ptosis<sup>30</sup>. The identification of such SVs demonstrates the power of LRS in disease diagnosis.

Overall, LRS methods offer a new era in clinical studies as they are becoming suitable as diagnostic tools for many complex diseases. They can detect a wide range of SVs, assemble genomes de novo, and do haplotype phasing. Furthermore, NGS techniques require additional tests because of their technical limitations, such as qPCR and Southern blot. These tests not only increase the turnaround time and consume more resources, but they also raise the overall cost of genome analysis. On the other hand, LRS methods do not require such additional tests. For this reason, LRS methods offer a faster and relatively cost-effective genome analysis approach in the long run.

## RESOLVING COMPLEX SMN LOCUS BY LRS METHODS

---

The first complete human genome was assembled in 2009 by the Genome Reference Consortium (GRC). In 2019, GRC has released its latest patch, GRCH38.p13<sup>61</sup>. The NGS methods used in this reference genome were only able to sequence 92% of the genome. The remaining 8%, which corresponds to 151 Mb, is considered camouflaged due to low read depth or low mapping quality<sup>62</sup>. Because of its complexity, the SMA-associated *SMN* locus is also considered a camouflaged region<sup>63,64</sup>.

With the development of LRS methods, the remaining 8% of the camouflaged regions became accessible for in-depth analysis<sup>63,65,66</sup>. LRS method PacBio HiFi has revealed 81 Mb camouflaged regions for the first time, including the entire *SMN* locus<sup>67</sup>. Besides copy number variances, the study has discovered functional changes in the organization and composition of *SMN2* sequence as well. Therefore, this study has shown complex loci such as *SMN* could be resolved by LRS methods.

NGS methods fail to resolve *SMN* locus for five reasons. One, *SMN* locus contains Alu repetitive sequences and pseudogenes within large inverted segmental duplications<sup>32,68,69</sup>. Two, the *SMN1* and *SMN2* sequences are homologous, which makes the separation of the two genes by NGS methods harder<sup>47</sup>. Three, the presence of *SMN2* CNVs suggests *SMN* locus is not stable and contains hotspots. These hotspots can undergo large rearrangements beyond the detection capability of NGS approaches<sup>32,70,71</sup>. Reason four is the gene hybridization events between *SMN1* and *SMN2*, resulting in mosaic sequences<sup>18,32,47</sup>. Lastly, the *SMN* locus contains genes besides *SMN* called *NAIP*, *GTF2H2A*, and *SERF1A*<sup>19</sup>. These genes also exhibit CNVs and SVs, adding to the complexity of the *SMN* locus<sup>32</sup>.

Previous studies have shown that gene conversion occurs within the *SMN* locus. However, its detection and explanation have long been a challenge due to the technological limitations of NGS. MLPA analysis and Illumina sequencing methods have determined the *SMN1*-*SMN2* gene hybridization events in 286 SMA patients<sup>27</sup>. Although they managed to identify hybrid genes, they have failed to explain the exact DNA sequences and mechanisms associated with SVs within these patients. Surprisingly, the study has found *SMN1*-*SMN2* gene hybridization events occur more frequently than previously believed. Genomic analysis of *SMN* locus across black and white populations has revealed that gene hybridization events might be responsible for the observed *SMN2* copy number disparities. White European descendants are determined to have relatively higher *SMN2* copies due to complete deletion or gene conversion events<sup>18</sup>. However, gene conversion events occur less frequently in black populations, where *SMN1* copy numbers are higher than *SMN2* numbers<sup>32</sup>. Identification of gene hybridization events holds clinical importance as

they could also potentially mask deletions within *SMN1*. For instance, non-SMA individuals with multiple *SMN1* exons 1-7 have been identified in black South African populations<sup>32</sup>. The additional copies in these individuals are believed to come from gene hybridization events. Although these individuals do not have the full *SMN1* sequence, traditional gene panels may register them as *SMN1* positive, hiding pathogenic CNVs and heterozygous *SMN1* deletions.

Furthermore, LRS methods could identify potential SVs to explain discordant SMA phenotypes. Genetic tests use the *SMN2* copy number of the patients as a parameter to make assumptions about the severity of the SMA phenotypes<sup>11,72</sup>. However, comparing *SMN2* copy numbers with SMA diagnosis has shown that *SMN2* is not a good predictor of symptom severity in SMA patients. The strongest correlation occurs between *SMN2* copy number and SMA symptoms in extreme cases. Patients with very high *SMN2* copy numbers (> 6) have moderate symptoms, whereas those with low *SMN2* copy numbers (< 2) have severe symptoms<sup>27</sup>. Yet, *SMN2* CNVs fail to explain the observed symptoms in a large portion of the SMA patients<sup>27</sup>. For instance, an individual with 3 *SMN2* copies is diagnosed with type 1c SMA, a diagnosis much severe than expected<sup>27</sup>. On the other hand, SMA patients with 2 *SMN2* copies and c.859G>C mutation are expected to show severe symptoms but exhibit much milder symptoms<sup>27</sup>. Furthermore, siblings with same *SMN2* copy numbers can show varying SMA symptoms<sup>27</sup>. These examples above show that *SMN2* copy number cannot fully explain the clinical variability of the SMA.

Intragenic mutations within *SMN2*, such as partial deletions and hybridization events, have a role in discordant SMA phenotypes (Wadman et al., 2020). Unfortunately, not all genomic reasons behind the clinical variability of SMA are known. When detecting gene hybridization and SVs, tests such as MLPA and NGS can produce misleading results due to their limitations. On the other hand, LRS methods, especially optical mapping, analyze DNA in detail and detect pathogenic variations. For instance, optical mapping is used in the detection of possibly pathogenic variations causing Poretti-Boltshauser syndrome (PBS). Surprisingly, patients also observed with enlarged ventricles in the heart, which is uncommon for this disease<sup>73</sup>. Bionano optical mapping analysis of genetic samples taken from parents whom two children are affected by PBS have revealed that besides commonly observed SNVs, children had an additional ~48 kb duplication within the *LAMA1* locus. Analysis of the parental genomes have revealed that PBS associated SNVs are inherited from the father, whereas, the duplication that responsible for the enlargement of the ventricles is inherited from the mother<sup>73</sup>. This study has shown that optical mapping technology can be used in the detection of SVs that contribute to phenotypical variations to broaden the PBS phenotype spectrum. By extensively analyzing genomic variants, LRS methods could also explain the discordant SMA

symptoms as well. As it can be seen from the PBS study, disease associated phenotypes can vary by the presence of duplications. If duplications of sequences that have not been detected by NGS methods before have a possible role in the SMA phenotype variability, LRS methods can be used in their detection to provide an explanation to observed discordant SMA phenotypes.

Since large deletions and duplications are very common within the *SMN* locus, alterations in *SMN1* copy numbers frequently occur in populations. In some cases, individuals can become heterozygous for *SMN1* through deletion on one chromosome and gain of copy on the other (2 + 0). Translocation of *SMN1* from one parental chromosome to the other can also lead to generation *SMN1* heterozygosity. Such individuals are not symptomatic but are silent carriers of SMA. Silent carriers have a greater risk of passing SMA to their progeny, especially if the other parent is also heterozygous for *SMN1* (1 + 0 or 2 + 0). Studies that compare *SMN2* CNVs between different populations have found black populations and Ashkenazi Jewish populations to have a higher ratio of *SMN2* silent carriers<sup>28,32,70</sup>. The CNV of *SMN2* is assessed by dosage-specific methods like MLPA and qPCR<sup>27,28,31,32,47</sup>. Yet, these methods cannot distinguish the silent carriers from wild-type individuals (1 + 1). As a result, *SMN* carrier screening tests especially give higher rates of false negatives in these populations<sup>28,32,47</sup>.

Separated parental chromosomes can be sequenced by ONT for haplotype phasing. Furthermore, ONT is useful in detection of large SVs. Because of these two reasons, ONT has been shown to be highly suitable for detection of translocations as well as CNV screening across parental chromosomes<sup>59,74</sup>. Individuals with genomic translocations usually appear normal but have a higher risk of infertility or miscarriage. ONT is used in genetic testing of parental chromosomes preimplantation to in vitro fertilization treatment to screen for individuals with genetic imbalances created by translocations<sup>74</sup>. In a different study, ONT is also used in the detection of SVs associated with Lynch syndrome. Nanopore sequencing has detected an 87 kb long pathogenic deletion as well as the breakpoints of SVs. This study has shown that nanopore sequencing could replace MLPA for the screening of SVs in hereditary disorders<sup>59</sup>. Therefore, LRS approaches can detect haplotype diversity in individuals and contribute to preventative actions by identifying and advising silent carrier parents.

## DISCUSSION

---

Human reference genome (HRG) is frequently used in NGS studies to provide reference sequences for accurate sequence alignment. Sequencing studies that use NGS to assemble genomes *de novo* also use HRG for verification. Due to these aforementioned reasons, HRG is accepted as the golden standard for

genome sequencing accuracy. However, recent telomere-to-telomere and whole genome sequencing studies using LRS methods have identified novel and more frequent SVs than found in HRG<sup>62,67</sup>. These findings indicate that human genome is much more diverse and as a single reference genome, HGR cannot represent the true genomic variety in populations. Thus, using a single reference genome can lead to misleading assumptions in the clinically relevant variations. For instance, CNVs are observed to be strongly underrepresented within the HRG<sup>67</sup>. As a result, a normally occurring CNV may be observed as genomic aberration when HRG is used as reference.

On the other hand, creating numerous reference genomes from diverse populations can provide a more realistic picture of human DNA's true genomic variation. To this end, various countries have already assembled their reference genomes. These countries include the Netherlands, the United Arab Emirates (UAE), China, South Africa, Iceland, and Denmark<sup>32,75-79</sup>. Although generating multiple human reference genomes from different populations is a step in the right direction, there is still more room for improvement. The UAE, for example, has created a main allele reference genome rather than a complete Arabic genome<sup>77</sup>. The genome of the Netherlands is also assembled from short-read sequences and contains gaps<sup>76</sup>. These projects can be improved by utilizing LRS to cover these mentioned gaps. Besides the incomplete reference genomes, the ability to combine and compare different reference genomes poses another challenge. Fortunately, a flow alignment method that allows users to integrate various population genomes to decrease possible reference bias has been developed<sup>80</sup>. Another advantage of establishing multiple reference genomes is that they can be used in the creation of comprehensive, clinically relevant SV libraries. Although many diseases are associated with SVs, not all SVs have pathogenic phenotypes<sup>2,58</sup>. As a result, pathogenic SVs can be prioritized in clinical research for accurate detection of disease-causing alterations.

In addition to generate SV libraries, LRS can also be used in the identification of clinically relevant genomic interactions between genes. By assembling whole-genome sequences, LRS platforms could identify relationships between genetic alterations outside of *SMN* locus and SMA phenotype variation. Such findings could especially provide an explanation to some of the discordant SMA phenotypes. For instance, *Plastin-3 (PLS3)* is found to have a role in SMA symptomatic presentations<sup>81</sup>. As an isoform of actin-binding proteins, *PLS3* functions in the stability and organization of F-actin filaments in cytoskeleton. High *PLS3* expression is shown to be correlated with milder SMA symptoms, which suggests that *PLS3* is used as a compensatory mechanism to *SMN* protein deficiency in humans<sup>81,82</sup>. In addition to *PLS3*, genes *UBE1*, *GARS* and *SETX* are also implicated in SMA development<sup>83</sup>. These findings suggest that SMA is a

highly complex disease, and its symptomatic outcomes may be determined by multiple genes. Because of their ability to assemble whole genome and detect complex genomic variations, LRS methods could be used in generating extensive patient profiles. These profiles could be used to elaborate a possible relationship between distinctive genomic alteration patterns and symptomatic outcomes of SMA.

Therefore, it should not be an exaggeration to suggest that LRS will likely become the next gold standard in genome sequencing in clinical biology. One of the advantages of LRS is its ability to function outside of the laboratory. All NGS sequencers are benchtop or industrial-sized machines that are not portable. Because of their bulky design, these devices can only be used in comprehensive laboratory environments. The ONT sequencer MinION, on the other hand, is a highly portable, handheld device. Its small size makes this sequencer very useful, especially for bedside analysis. Indeed, MinION has been used to sequence genomes in various places to demonstrate its functionality. Some of the extreme environments include remote villages in West Africa, the Arctic, and the International Space Station<sup>84-86</sup>.

Another key factor that makes LRS methods popular is their user-friendly, easy-to-navigate workflows. NGS methods require highly skilled technicians to conduct delicate processes to ensure that generated libraries are not contaminated and complete. On the other hand, sample preparation for LRS platforms is relatively less tedious and easy to follow<sup>87</sup>. Furthermore, compared to NGS data analysis tools, LRS platforms have developed programs such as EPI2ME with built-in algorithms that can conduct complex data analysis on sequencing files. Therefore, even individuals that do not have a strong bioinformatics background can still perform their own data analysis<sup>88</sup>. These relatively simpler workflows and tools make the whole sequencing process faster, which is a trait that is always been sought in genome sequencing.

In addition to their functionality, LRS platforms are also becoming more cost-effective. Although the current cost of LRS is higher than NGS, the adaptation of high-throughput workflows in these systems has made the overall cost decrease. For instance, PacBio has developed multiplex sequencing workflows to enable large numbers of sample libraries to be pooled and sequenced simultaneously on an SMRT Cell. By combining multiple libraries, more samples can be tested in a single run, thereby decreasing the overall cost per sample<sup>3,89</sup>. It should also be worth mentioning that MinION by ONT is the cheapest sequencer on the market. The overall cost of this device is a little over \$1000, which makes this sequencer a budget-friendly option for laboratories that have limited financial power but are interested in genome sequencing<sup>90</sup>. Given these points, LRS has the potential to become the future of genetic sequencing.

## CONCLUSION

---

As the cost of genome sequencing gets lower, and algorithmic tools evolve, LRS methods are expected to become the future of genome sequencing. As it is established in this paper, LRS methods can overcome the limitations of NGS techniques for accurately detecting SVs and CNVs. Furthermore, they can bypass the need to use qPCR, MLPA, etc. for further validation of their results. Therefore, LRS methods offer a faster and cheaper way to sequence DNA. Since no additional tests are needed with LRS, these platforms also become relatively cost-effective in the long term for DNA analysis. Hence, LRS methods have the potential to be used in laboratory techniques in clinical studies for diagnosis, prognosis, and targeted treatment. The *SMN* locus is one of these key areas within the genome that can be resolved by LRS methods. Because of its complex structure, sequencing the *SMN* locus with NGS methods has been a clinical challenge. However, LRS methods have proven to be capable of sequencing complex regions. Their use in *SMN* locus can especially provide new insights on the phenotypical variety of SMA and genetic alteration patterns. Moreover, LRS methods can replace traditionally used tests, such as MLPA to determine *SMN2* CNVs. This way, silent carriers can be detected more effectively, and the necessary genetic counseling can be provided to the individuals that have a higher risk of giving birth to children with SMA.

## REFERENCES

---

1. Sanger, F. *et al.* Nucleotide sequence of bacteriophage  $\phi$ X174 DNA. *Nature* **265**, (1977).
2. Kumar, K. R., Cowley, M. J. & Davis, R. L. Next-Generation Sequencing and Emerging Technologies. *Seminars in Thrombosis and Hemostasis* **45**, 661–673 (2019).
3. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology* **122**, (2018).
4. Morozova, O. & Marra, M. A. Applications of next-generation sequencing technologies in functional genomics. *Genomics* **92**, (2008).
5. Yohe, S. & Thyagarajan, B. Review of clinical next-generation sequencing. *Archives of Pathology and Laboratory Medicine* vol. 141 1544–1557 (2017).
6. Zhong, Y., Xu, F., Wu, J., Schubert, J. & Li, M. M. Application of Next Generation Sequencing in Laboratory Medicine. *Annals of Laboratory Medicine* **41**, 25–43 (2020).
7. van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends in Genetics* **34**, (2018).

8. Payne, A., Holmes, N., Rakyar, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, (2019).
9. Lam, E. T. *et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology* **30**, 771–776 (2012).
10. Hendrickson, B. C. *et al.* Differences in SMN1 allele frequencies among ethnic groups within North America. *Journal of Medical Genetics* **46**, 641–644 (2009).
11. Lefebvre, S. *et al.* Identification and Characterization of a Spinal Muscular Atrophy-Determining Gene. *Cell* vol. 80 (1995).
12. Bussaglia, E. *et al.* A frame-shift deletion in the survival motor neuron gene in Spanish spinal muscular atrophy patients. <http://www.nature.com/naturegenetics> (1995).
13. Devriendt, K. *et al.* Clinical and molecular genetic features of congenital spinal muscular atrophy. *Annals of Neurology* **40**, (1996).
14. Hahnen, E. *et al.* Molecular analysis of candidate genes on chromosome 5q13 in autosomal recessive spinal muscular atrophy: evidence of homozygous deletions of the SMN gene in unaffected individuals. *Human Molecular Genetics* **4**, (1995).
15. Hahnen, E., Schönling, J., Rudnik-Schöneborn, S., Zerres, K. & Wirth, B. Hybrid survival motor neuron genes in patients with autosomal recessive spinal muscular atrophy: new insights into molecular mechanisms responsible for the disease. *American journal of human genetics* **59**, (1996).
16. Matthijs, G. *et al.* Unusual molecular findings in autosomal recessive spinal muscular atrophy. *Journal of Medical Genetics* **33**, (1996).
17. Rodrigues, NandaR., Campbell, L., Owen, N., Rodeck, CharlesH. & Davies, KayE. Prenatal diagnosis of spinal muscular atrophy by gene deletion analysis. *The Lancet* **345**, (1995).
18. van der Steege, G. *et al.* Apparent gene conversions involving the SMN gene in the region of the spinal muscular atrophy locus on chromosome 5. *American journal of human genetics* **59**, (1996).
19. Butchbach, M. E. R. Genomic variability in the survival motor neuron genes (Smn1 and smn2): Implications for spinal muscular atrophy phenotype and therapeutics development. *International Journal of Molecular Sciences* **22**, (2021).
20. Wirth, B. *et al.* Mapping of the Spinal Muscular Atrophy (SMA) Gene to a 750-kb Interval Flanked by Two New Microsatellites. *European Journal of Human Genetics* **3**, (1995).
21. Parsons, D. W. *et al.* Intragenic telSMN Mutations: Frequency, Distribution, Evidence of a Founder Effect, and Modification of the Spinal Muscular Atrophy Phenotype by cenSMN Copy Number. *The American Journal of Human Genetics* **63**, (1998).
22. Blasco-Pérez, L. *et al.* Beyond copy number: A new, rapid, and versatile method for sequencing the entire SMN2 gene in SMA patients. *Human Mutation* **42**, (2021).

23. Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. *A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. Genetics* vol. 96 www.pnas.org. (1999).
24. Singh, N. K., Singh, N. N., Androphy, E. J. & Singh, R. N. Splicing of a Critical Exon of Human *Survival Motor Neuron* Is Regulated by a Unique Silencer Element Located in the Last Intron. *Molecular and Cellular Biology* **26**, (2006).
25. Kolb, S. J. & Kissel, J. T. Spinal Muscular Atrophy. *Neurologic Clinics* **33**, (2015).
26. Lopez-Lopez, D. *et al.* SMN1 copy-number and sequence variant analysis from next-generation sequencing data. *Human Mutation* **41**, (2020).
27. Wadman, R. I. *et al.* Intragenic and structural variation in the SMN locus and clinical variability in spinal muscular atrophy. *Brain Communications* **2**, (2020).
28. Luo, M. *et al.* An Ashkenazi Jewish SMN1 haplotype specific to duplication alleles improves pan-ethnic carrier screening for spinal muscular atrophy. *Genetics in Medicine* **16**, (2014).
29. Bowerman, M. *et al.* Therapeutic strategies for spinal muscular atrophy: SMN and beyond. *DMM Disease Models and Mechanisms* vol. 10 943–954 (2017).
30. Mizuguchi, T. *et al.* Pathogenic 12-kb copy-neutral inversion in syndromic intellectual disability identified by high-fidelity long-read sequencing. *Genomics* **113**, (2021).
31. Alías, L. *et al.* Improving detection and genetic counseling in carriers of spinal muscular atrophy with two copies of the *SMN1* gene. *Clinical Genetics* **85**, (2014).
32. Vorster, E., Essop, F. B., Rodda, J. L. & Krause, A. Spinal Muscular Atrophy in the Black South African Population: A Matter of Rearrangement? *Frontiers in Genetics* **11**, (2020).
33. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *Journal of Human Genetics* **59**, (2014).
34. Majewski, J., Schwartzenruber, J., Lalonde, E., Montpetit, A. & Jado, N. What can exome sequencing do for you? *Journal of Medical Genetics* **48**, (2011).
35. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* **30**, (2014).
36. Chorev, M. & Carmel, L. The Function of Introns. *Frontiers in Genetics* **3**, (2012).
37. Jackson, S. A., Cannone, J. J., Lee, J. C., Gutell, R. R. & Woodson, S. A. Distribution of rRNA Introns in the Three-dimensional Structure of the Ribosome. *Journal of Molecular Biology* **323**, (2002).
38. Su, Y. *et al.* Deciphering Neurodegenerative Diseases Using Long-Read Sequencing. *Neurology* vol. 97 423–433 (2021).
39. Ardui, S., Ameer, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* **46**, (2018).

40. Ebbert, M. T. W. *et al.* Long-read sequencing across the C9orf72 “GGGGCC” repeat expansion: Implications for clinical use and genetic discovery efforts in human disease. *Molecular Neurodegeneration* **13**, (2018).
41. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155–1162 (2019).
42. de Roeck, A. *et al.* NanoSatellite: Accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. *Genome Biology* **20**, (2019).
43. Chan, S. *et al.* Structural Variation Detection and Analysis Using Bionano Optical Mapping. in (2018). doi:10.1007/978-1-4939-8666-8\_16.
44. Goldrich, D. Y. *et al.* Identification of Somatic Structural Variants in Solid Tumors by Optical Genome Mapping. *Journal of Personalized Medicine* **11**, (2021).
45. Bocklandt, S., Hastie, A. & Cao, H. Bionano Genome Mapping: High-Throughput, Ultra-Long Molecule Genome Analysis System for Precision Genome Assembly and Haploid-Resolved Structural Variation Discovery. in (2019). doi:10.1007/978-981-13-6037-4\_7.
46. Watson, C. M. *et al.* Cas9-based enrichment and single-molecule sequencing for precise characterization of genomic duplications. *Laboratory Investigation* **100**, 135–146 (2020).
47. Chen, X. *et al.* Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genetics in Medicine* **22**, (2020).
48. Mailman, M. *et al.* Hybrids monosomal for human chromosome 5 reveal the presence of a spinal muscular atrophy (SMA) carrier with two SMN1 copies on one chromosome. *Human Genetics* **108**, (2001).
49. Anhuf, D., Eggermann, T., Rudnik-Schöneborn, S. & Zerres, K. Determination of SMN1 and SMN2 copy number using TaqMan™ technology. *Human Mutation* **22**, (2003).
50. Martignano, F. *et al.* Nanopore sequencing from liquid biopsy: analysis of copy number variations from cell-free DNA of lung cancer patients. *Molecular Cancer* **20**, (2021).
51. Brookes, K. J. The VNTR in complex disorders: The forgotten polymorphisms? A functional way forward? *Genomics* **101**, (2013).
52. Paulson, H. Repeat expansion diseases. in (2018). doi:10.1016/B978-0-444-63233-3.00009-9.
53. Höijer, I. *et al.* Detailed analysis of *HTT* repeat elements in human blood using targeted amplification-free long-read sequencing. *Human Mutation* **39**, (2018).
54. Ishiura, H. *et al.* Noncoding CGG repeat expansions in neuronal intranuclear inclusion disease, oculopharyngodistal myopathy and an overlapping disease. *Nature Genetics* **51**, (2019).

55. Loomis, E. W. *et al.* Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Research* **23**, (2013).
56. McFarland, K. N. *et al.* Paradoxical effects of repeat interruptions on spinocerebellar ataxia type 10 expansions and repeat instability. *European Journal of Human Genetics* **21**, (2013).
57. Sone, J. *et al.* Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nature Genetics* **51**, (2019).
58. de Coster, W. *et al.* Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research* **29**, 1178–1187 (2019).
59. Yamaguchi, K. *et al.* Application of targeted nanopore sequencing for the screening and determination of structural variants in patients with Lynch syndrome. *Journal of Human Genetics* **66**, (2021).
60. Cope, H. *et al.* Detection of a mosaic *CDKL5* deletion and inversion by optical genome mapping ends an exhaustive diagnostic odyssey. *Molecular Genetics & Genomic Medicine* **9**, (2021).
61. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27**, (2017).
62. Nurk, S. *et al.* The complete sequence of a human genome Affiliations are listed at the end \* Equal contribution †. doi:10.1101/2021.05.26.445798.
63. Ebbert, M. T. W. *et al.* Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology* **20**, (2019).
64. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431**, (2004).
65. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, (2021).
66. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, (2020).
67. Vollger, M. R. *et al.* SEGMENTAL DUPLICATIONS AND THEIR VARIATION IN A COMPLETE HUMAN GENOME. doi:10.1101/2021.05.26.445678.
68. Bürglen, L. *et al.* Structure and Organization of the Human Survival Motor Neurone (SMN) Gene. *Genomics* **32**, (1996).
69. Selig, S. *et al.* Expressed cadherin pseudogenes are localized to the critical region of the spinal muscular atrophy gene. *Proceedings of the National Academy of Sciences* **92**, (1995).
70. Chimusa, E. R. *et al.* A Genomic Portrait of Haplotype Diversity and Signatures of Selection in Indigenous Southern African Populations. *PLOS Genetics* **11**, (2015).

71. Choudhury, A. *et al.* Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC Genomics* **15**, (2014).
72. Gennarelli, M. *et al.* Survival Motor-Neuron Gene Transcript Analysis in Muscles from Spinal Muscular-Atrophy Patients. *Biochemical and Biophysical Research Communications* **213**, (1995).
73. Chen, M. *et al.* Identification of a likely pathogenic structural variation in the LAMA1 gene by Bionano optical mapping. *npj Genomic Medicine* **5**, (2020).
74. Chow, J. F. C., Cheng, H. H. Y., Lau, E. Y. L., Yeung, W. S. B. & Ng, E. H. Y. Distinguishing between carrier and noncarrier embryos with the use of long-read sequencing in preimplantation genetic testing for reciprocal translocations. *Genomics* **112**, (2020).
75. Beyter, D. *et al.* Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nature Genetics* **53**, 779–786 (2021).
76. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *European journal of human genetics : EJHG* **22**, (2014).
77. Daw Elbait, G., Henschel, A., Tay, G. K. & al Safar, H. S. A Population-Specific Major Allele Reference Genome From The United Arab Emirates Population. *Frontiers in Genetics* **12**, (2021).
78. Huang, C. *et al.* An integrated Asian human SNV and indel benchmark established using multiple sequencing methods. *Scientific Reports* **10**, (2020).
79. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* **548**, (2017).
80. Chen, N.-C., Solomon, B., Mun, T., Iyer, S. & Langmead, B. Reference flow: reducing reference bias using multiple population genomes. *Genome Biology* **22**, (2021).
81. Oprea, G. E. *et al.* Plastin 3 Is a Protective Modifier of Autosomal Recessive Spinal Muscular Atrophy. *Science* **320**, (2008).
82. Yanyan, C. *et al.* Correlation of PLS3 expression with disease severity in children with spinal muscular atrophy. *Journal of Human Genetics* **59**, (2014).
83. Šoltić, D. & Fuller, H. R. Molecular Crosstalk Between Non-SMN-Related and SMN-Related Spinal Muscular Atrophy. *Neuroscience Insights* **15**, (2020).
84. Castro-Wallace, S. L. *et al.* Nanopore DNA Sequencing and Genome Assembly on the International Space Station. *Scientific Reports* **7**, (2017).
85. Goordial, J. *et al.* In Situ Field Sequencing and Life Detection in Remote (79°26'N) Canadian High Arctic Permafrost Ice Wedge Microbial Communities. *Frontiers in Microbiology* **8**, (2017).

86. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, (2016).
87. Lin, B., Hui, J. & Mao, H. Nanopore technology and its applications in gene sequencing. *Biosensors* vol. 11 (2021).
88. Petersen, L. M., Martin, I. W., Moschetti, W. E., Kershaw, C. M. & Tsongalis, G. J. *Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing*. <https://doi.org/10> (2019).
89. Qian, J. *et al.* Multiplexed Non-barcoded Long-Read Sequencing and Assembling Genomes of Bacillus Strains in Error-Free Simulations. *Current Microbiology* **77**, (2020).

## APPENDIX A

	<b>Instrument</b>	<b>Maximum throughput</b>	<b>Maximum output of reads</b>	<b>Read length</b>	<b>Sequencing run time (in hours)</b>	<b>Cost</b>	<b>Main applications</b>
Illumina	ISEq	1.2 Gb*	4 million	2 x 150 bp	9.5 - 19	\$25 - \$115 <sup>a</sup>	Small whole-genome sequencing, Targeted gene sequencing
	MiniSeq	7.5 Gb	25 million	2 x 150 bp	4 - 24	\$64 <sup>a</sup>	Small whole-genome sequencing, Targeted gene sequencing, Targeted gene expression profiling, 16S metagenomic sequencing
	MiSeq	15 Gb	25 million	2 x 300 bp	4 - 55	\$159 - \$318 <sup>a</sup>	Small whole-genome sequencing, Targeted gene sequencing, 16S metagenomic sequencing
	NextSeq 500	120 Gb	400 million	2 x 150 bp	12 - 30	\$285 <sup>a</sup>	Small whole-genome sequencing, Targeted gene sequencing, Transcriptome sequencing
	NextSeq 1000 & 2000	360 Gb	1.2 billion	2 x 150 bp	11 - 48	\$450 <sup>a</sup>	Small whole-genome sequencing, Exome & large panel sequencing, Single-cell profiling, Transcriptome sequencing, miRNA & small RNA analysis
	NovaSeq	6000 Gb	20 billion	2 x 250 bp	13 - 44	\$4,500 - \$5,400 <sup>a</sup>	Large whole-genome sequencing, Exome & large panel sequencing, Single-cell profiling, Transcriptome sequencing, Chromatin analysis, Methylation sequencing, Metagenomic profiling, Cell-Free sequencing
Ion Torrent	Ion 510	0.3 – 1 Gb	2 – 3 million	200 - 400 bp	3 - 4.5	\$142 <sup>b</sup>	Targeted gene sequencing, Transcriptome sequencing, Small whole-genome sequencing
	Ion 520	0.6 – 2 Gb	3 – 6 million	200 - 600 bp	3 - 12	\$131 <sup>b</sup>	
	Ion 530	1.5 – 8 Gb	9 – 20 million	200 - 600 bp	4 - 21	\$102 <sup>b</sup>	
	Ion 540	10 – 30 Gb	60 – 80 million	200 bp	6.5 - 20	\$219 <sup>c</sup>	
	Ion 550	25 – 50 Gb	100 – 130 million	200 bp	8.5 - 12	\$401 <sup>c</sup>	

## Appendix A (Continued)

	Instrument	Maximum throughput	Maximum output of reads	Read length	Sequencing run time (in hours)	Cost	Main applications
Nanoball	DNBSEQ – T7	6 Tb**	5000 million	150 bp	24 – 30	≥ \$600 <sup>d</sup>	Whole-genome sequencing, Deep exome sequencing, Transcriptome sequencing, Targeted panel project
	DNBSEQ – G400	1440 Gb	1500 - 1800 million	50 - 300 bp	13 - 109		Whole-genome sequencing, Whole-exome sequencing, Transcriptome sequencing
	DNBSEQ – G50	150 Gb	100 - 500 million	50 – 100 bp	9 - 40		Small whole-genome sequencing, Targeted sequencing, Low-pass whole-genome sequencing
PacBio	Sequel	20 Gb	500,000	15 kb	20	\$250 <sup>e</sup>	Whole-genome sequencing, Targeted sequencing, Epigenetics, Population sequencing, RNA sequencing
	Sequel II	30 Gb per SMRT cell	4 million	15 kb	30	\$1950 <sup>f</sup>	
	Sequel IIe	30 Gb per SMRT cell	4 million	20 - 30 kb	30	\$1950 <sup>f</sup>	
ONT	MinION	50 Gb	0.5 million	Hundreds of kb to several Mb	72	\$900 <sup>g</sup>	Whole-genome sequencing, Targeted sequencing, Epigenetics, RNA sequencing, Gene expression analysis
	GridION	50 Gb per flow cell	2.5 million		72	\$900 <sup>g</sup>	
	PromethION	290 Gb per flow cell	375 million		64	\$625-\$2000 <sup>g</sup>	
Bionano	Saphyr	480 Gb	NA	150 kb – 3 Mb	< 36	\$550 <sup>h</sup>	<i>De novo</i> genome assembly, Detection of structural variations (> 500 bp)

Table 1 - Detailed list of NGS and LRS sequencers and their specifications.

<sup>a</sup> Price acquired from <https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-cost.html><sup>b</sup> Price per sample (acquired from <https://www.thermofisher.com/us/en/home/order.html>)<sup>c</sup> Price per reaction (acquired from <https://www.thermofisher.com/us/en/home/order.html>)<sup>d</sup> Cost of human whole genome sequencing (acquired from <https://www.bgi.com/us/sequencing-services/dna-sequencing/whole-genome-sequencing/>)<sup>e</sup> High coverage long read amplicon sequencing (acquired from [www.mrdnalab.com](http://www.mrdnalab.com))<sup>f</sup> Price acquired from University of Washing PacBio Sequencing Services (<https://pacbio.gs.washington.edu/>)<sup>g</sup> Price per Gb/per flow cell (acquired from <https://nanoporetech.com/products/comparison>)<sup>h</sup> Price per genome (acquired from <https://bionanogenomics.com/products/bionano-data-options/>)

\* Gb = Gigabase \*\* Tb = Terabase