Introduction

Structural variants (SVs) are polymorphisms involving a segment of DNA that differs between individuals. In cancer, these differ between somatic and normal samples. SVs include genomic rearrangements such as deletions, insertions, inversions, duplications, and translocations whose sizes are larger than 50 bp. Research into SVs have established their importance to molecular biology, elucidating their role in various diseases and regulation of gene expression.

However, detecting SVs poses challenges as the evidence for SVs often resembles common sequencing and alignment artifacts. Traditional sequencing approaches have relied on short read sequencing (with read lengths between 36 bp ~ 400 bp). However, detecting SVs with short read sequencing is more complicated as short read data accuracy can be impacted by a number of issues, including:

- genomes with repetitive sequences where SVs tend to occur
- low read coverages in GC-poor or GC-rich regions
- genome sequencing bias introduced by PCR amplification during library construction

The advent of new long-read sequencing technologies by PacBio and Oxford Nanopore sequencing has led to a revolution in genome sequencing, where long reads up to 100 Kb can be sequenced in a single run without PCR amplification. Long read sequencing approaches provide the opportunity to more accurately and reliably detect SVs at a much higher resolution.

Methodology

**Library Construction and Sequencing**

**Oxford Nanopore**

Using qualified high-quality DNA and the Oxford Nanopore library construction kit:
1. Size-Selection with the BluePippin/Pippin HT System/Sage HLS
2. Library Construction
   1) DNA Damage repair, End Repair, A-Tailing and clean-up
   2) Adapter Ligation
3) Magnetic Beads Purification and Qubit Quantification
4. Priming and loading the flow cell
5. On-board sequencing of the library

**PacBio**

Preparation of library DNA: high-quality whole genome long fragment DNA
1. gDNA shearing and clean-up (HiFi)
2. Remove Single-Strand Overhangs
3. DNA Damage repair, End Repair, A-Tailing
4. Adapter connection: the double-stranded positive and negative chains to obtain a dumbbell-like (“Tortle horse Ring”) structure, called SMRT Bell
5. Size-Selection with the Sage ELF/BluePippin/Pippin HT System
6. Anneal and Bind SMRTbell Library
7. Prepare for Sequencing
Long-read Human Whole Genome Resequencing

Sequencing

Oxford Nanopore Sequencing

All Oxford Nanopore Technology (ONT) sequencing devices use flow cells which contain an array of tiny holes — nanopores — embedded in an electro-resistant membrane. Each nanopore corresponds to its own electrode connected to a channel and sensor chip, which measures the electric current that flows through the nanopore. When a molecule passes through a nanopore, the current is disrupted to produce a characteristic ‘squiggle’. The squiggle is then decoded using base calling algorithms to determine the DNA or RNA sequence in real time. Read more about how Nanopore sequencing works on the ONT website.

PacBio sequencing

SMRT (Single-Molecule, Real-Time) sequencing is single-molecule real-time sequencing (sequencing-by-synthesis). Fluorescently labeled dNTPs randomly enter the detection area through Brownian motion and bind to the polymerase. The bases that match the template form chemical bonds for much longer than the other bases. Therefore, the existence time of the fluorescent signal can be counted to distinguish matched bases from free bases. The DNA template sequence can be determined by counting the relationship between the four fluorescence signals and time [1]. Read more on the PacBio website.

Data Performance

The standard sample HG002, which is currently considered to be the most thoroughly studied diploid human genome in the world, and a high-confidence variant set, have both been released. These can be used as an important benchmark to understand the performance of sequence platforms and test results.

Basic information of test data

For the standard sample HG002, we tested 5 types of data of PB CLR, PB hifi, ONT R9, ONT Q20 and short reads with a sequencing depth of 30X (where PB CLR is the downloaded data from the official website, and the others are our sequencing data). The length distribution is as follows:
Based on the 4 types of long read data, we tested 3 different detection software’s respectively. From the test results, we can see that the data of ONT R9 based on cuteSV software has the best performance. The f1 value for long-read platform is far higher than the short-read data.

**SV detection accuracy and sensitivity**

Based on the 4 types of long read data, we tested 3 different detection software’s respectively. From the test results, we can see that the data of ONT R9 based on cuteSV software has the best performance. The f1 value for long-read platform is far higher than the short-read data.

**Variation detection based on ONT R9 30X SV compared with WGS SV**

The detection data of different variant types of several software’s are much larger than short-read sequencing. The length of each variant type is also much longer than short-read. Especially in the detection ability of SVs greater than 100bp, the long-read advantage is much more obvious.
Spinal Muscular Atrophy (SMA) is an autosomal recessive neuromuscular disease. A common motor neuron disease mainly characterized by muscle weakness contraction, low muscle tone, and weakened sputum response. There is no effective treatment yet. At present, genetics has been confirmed that SMA is closely related to two highly homologous genes, SMN1 and SMN2. The two genes are distinguished on exon 7 and exon 8. There is only a five-base difference for SMN1 and SMN2 throughout the DNA level and two-base in the coding region.

The development of long-read sequencing enables the analysis of regions which can be difficult for regular WGS. On the left is the mapping result to SMN1 gene whose mutations are responsible for the genetic disorder SMA and its homologue SMN2 gene on the right. These two genes are extremely similar, resulting in ambiguous mapping of short reads. This makes it impossible to analyze, but long-read sequencing successfully rescued those reads and properly mapped them using co-bar-coding information. This makes variant calling on highly homologous regions possible.

*The above analysis results are obtained from the BGI information analysis process. The test results do not represent delivery indicators. The final interpretation right belongs to Shenzhen BGI Co., Ltd.

References