**\section{Single molecule sequencing}**

\textit{There is nothing more powerful than an idea whose time has come}

\begin{flushright} [Victor Hugo] \end{flushright}

Single molecule sequencing technologies from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) is spearheading the next decade of genomics revolution. These upcoming technologies promises a new era of genomics where: 1) lower input material is required for library preparation and sequencing, 2) library preparation is location agnostic and does not require skilled technicians, 3) sequencing takes hours and not days, 4) higher base accuracy, 5) longer read length (10kb – 100kb), 6) simultaneous detection of genetic variations and base modifications and 7) nucleotide-resolution identification of structural variations. Despite these promising capabilities, the higher error rate and marginally longer read length of first generation of ONT reads and PacBio continuous long reads (CLR) limited their adoption. Illumina is still the primary sequencing method in most labs as per base sequencing cost is still cheaper with the Illumina platform. After decades of development, ONT and PacBio have introduced new sequencing instruments and library preparation that exceeds the capabilities of Illumina platform in read length and accuracy, enabling researchers to access the inaccessible regions of the genome and to explore biological phenomena that could not be explored before.

**\subsection{Nanopore sequencing}**

Cells use membrane proteins to move ions and molecules, critical to the maintenance of cellular function, across the permeable plasma membrane through passive and active transport \cite{}. David Deamer and George Church independently hypothesised 197X that a single strand of DNA molecule could be passed through a protein pore if voltage is applied through the membrane holding protein pore cite{}. If electrostatic potential is present across the protein. The disruption of the passage of ionic currents by the passage of the DNA molecule through the pore can be recorded and can be interpreted as a specific nucleotide base. Nanopore based sequencing methods promised 1) minimal library preparation, 2) ultra-fast native DNA and RNA sequencing and 3) unlimited read length

Today, Oxford Nanopore Technologies (ONT) has fulfilled many of these promises. To fulfil these promises, Deamer and colleagues had to demonstrate the potential of the Nanopore sequencing method through successive demonstrations and improvements of the technology that first shows that passage of the DNA through a pore and disruption of the ionic current is a detectable event \cite{}, and that a single nucleotide difference can be detected from a background of homopolymer sequence \cite{}. In addition, the first generation of pores based on $alpha$ had a pore that was too long such that 10-15 nucleotides will be interpreted as a single signal and hence, a pore that had a similar aperture, but shorter pore was required to improve the signal-to-noise ratio. MytA protein, hence, thereafter, was used for nanopore sequencing to improve the signal-to-noise ratio. To improve their base accuracy, ONT introduced 2D reads where both forward and reverser strand of the double-stranded DNA with a hairpin adapter could be sequenced through the nanopore \cite{}. ONT, however, long no supports 2D reads as a result of legal dispute with PacBio \cite{}.

ONT licensed these patents to commercialise the technology in 2005 and the most recent ONT reads are reported to have Q20 read accuracy \cite{}. To date, ONT reads have been successfully used to identify and characterise complex pathogenic mutations\cite{}, accelerate clinical diagnosis \cite{}, and to help the assembly of the complex regions in the human reference genome \cite{}. It could be said that ONT sequencing has fulfilled all of it promises and more.

**\subsection{Pacific Biosciences Single-Molecule Real-Time sequencing: the end}**

PacBio was founded in 2004 with aspirations to commercialise single molecule real time (SMRT) sequencing technology developed at Cornell University. The SMRT platform is the culmination of multiple technical innovations from a range of disciplines. The zero-mode-waveguide (ZMW), a nano-scaled hole fabricated in a metal film, for example, is at the heart of the SMRT platform. The ZMW acts as the sequencing unit and its unique properties help the SMRT platform achieve the high signal-to-noise ratio required to observe activity of individual DNA polymerases (DNAP)\cite{Levene2003-og}.

The metal film with the ZMW is placed on top of a glass and DNAP is immobilised at the bottom glass surface through surface chemistry modifications that prevents the adsorption of DNAP to the metal side walls\cite{Korlach2008-aq, Eid2009-ol}. A topologically circular template, also known as a SMRTbell template, is created through the attachment of hairpin adapters to a double-stranded DNA molecule (Figure X). The successful loading of SMRTbell template into a ZMW follows a Poisson distribution and typically 30 to 50\% of the ZMWs are classified as productive ZMWs where a single DNAP successfully initiates and completes rolling circle amplification. SMRT sequencing initially used $\Phi$29 DNAP for its high processivity, minimal amplification bias and ability to perform strand displacement DNA synthesis \cite{Eid2009-ol}. In addition, $\Phi$29 DNAP was engineered through site-directed mutagenesis to use fluorophore-labeled deoxyribonucleoside triphosphate (dNTP) during DNA elongation \cite{Korlach2008-fv,Eid2009-ol}.

Upon successful loading of SMRTbell template, free nucleotides are released above the ZMW array and free nucleotides diffuses in and out of the ZMW. DNAP binds and incorporates the correct nucleotide into the growing DNA strand, and upon nucleotide incorporation, DNAP cleaves the fluorophore from the nucleotide such that the synthesised DNA molecule consists of native DNA molecules. DNAP continues DNA elongation until DNA replication is terminated. The length of the reaction time is dependent on DNAP processivity and the presence of bulky DNA damage on the template DNA that can lead to premature termination of replication\cite{}. Illumination from the laser below the glass surface excites the fluorophore and the emitted fluorescence is measured. Image processor leverages the temporal difference between diffusion of free nucleotides (which occurs in microseconds) and nucleotide incorporation (which occurs in milliseconds) to separate the background fluorescence from free nucleotides and fluorescence from nucleotide bound to DNAP. In addition, the size and shape of the ZMW prevents laser light from passing through the ZMW and limits the illumination to the bottom of the ZMW, which further increases the signal-to-noise ratio. As the four dNTPs are each labelled with a different fluorophore, each nucleotide can be identified from their unique fluorescence\cite{Eid2009-ol}. DNA base modification detection can also be achieved from analyzing DNAP kinetics, which is comprised of duration of fluorescence pulse, known as pulse width, and the duration between successive fluorescence pulses, referred to as interpulse duration \cite{Flusberg2010-ub}. To date, DNAP kinetics has been used to detect including base modifications such as N6-methyladenine, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine \cite{Flusberg2010-ub} and DNA damage such as O6-mmethylguanine, 1-methyladenine, O4-methylthymien, 5-hydroxycyostine, 5hydroxyuracil, 5-hydroxymethyluyracil and thymine dimers \cite{Clark2011-jz}.

SMRT platform capability was initially limited to continuous long read (CLR) generation with 10-15\% error rate \cite{Eid2009-ol} instead of circular consensus sequence (CCS) generation with 0.1-1\% error rate \cite{Wenger2019-pw}. This was because there is an inherent trade-off between read length and read accuracy while DNAP processivity is held as a constant. The earlier generations of DNAP had insufficient processivity to sequence both the forward and reverse strand of a SMRTbell template multiple times. In contrast, the more recent generations of DNAP have sufficient processivity to sequence the forward and reverse strand of long SMRTbell templates (>10kb) multiple times such that both long and accurate reads are produced \cite{Wenger2019-pw}. SMRT platform, hence, leveraged the improvements in DNAP processivity to first increase read length and subsequently, improve read accuracy.

The PacBio RS instrument with the first generation of polymerase and chemistry (P1-C1) produced continuous long reads (CLR) with an average read length of 1,500 bp with 10-15\% error rate \cite{}. In contrast, the most recent PacBio Revio instrument generates circular consensus sequence (CCS) reads with an average read length of 20,000 bp with 0.1-1\% error rate \cite{}. In addition, the PacBio RS instrument used the first generation of SMRTcell with 150,000 ZMWs \cite{} while the PacBio Revio instrument uses the latest SMRTcell with 25 million ZMWs, increasing the sequence throughput exponentially from 22 million bases to 90 billion bases per SMRTcell \cite{} (Figure X). Compared to Illumina sequencing, CLR sequencing had high error rate, higher cost per base and only marginal increases in read length. In addition, the shortage of bioinformatics algorithms to effectively process CLR reads with high error rate also slowed market adoption. The PacBio Revio instrument, however, can generate 30-fold CCS sequence coverage of the human genome under \$1000. The sequence data from a single SMRTcell, therefore, can be used for not only \textit{de novo} assembly \cite{} but also haplotype phased base modification\cite{}, SNP and indel, \cite{} and structural variation detection \cite{}, enabling the most comprehensive characterisation of both genetic and epigenetic variation from a single human individual. We also expect the sequence throughput per SMRTcell to increase exponentially in the foreseeable future with improvements in DNA processivity that increases CCS read length and advances in semiconductor fabrication technologies that doubles or triples the number of ZMWs per SMRTcell.

**\subsection{Long-read sequencing applications}**

In the beginning, long reads from ONT and PacBio SMRT platform did not have a competitive advantage compared to short reads from Illumina platform; Long reads were only marginally longer than short reads and their higher error rate made germline mutation detection more challenging. Long-read sequencing, most importantly, could not compete with short-read sequencing on sequencing cost.

A substantial increase in read length from ~1,500 bp to 10,000 bp with the introduction of XX chemistry for ONT and P5-C3 chemistry for PacBio Sequel I instrument reignited interest for new \textit{de novo} assembly algorithm development, full-length transcript sequencing and accessing the inaccessible regions of the genome.

**\subsubsection{\textit{De novo} assembly}**

Genomes are peppered with repetitive sequences. These repetitive sequences, for example, account for approximately 50\% of the human genome\cite{Lander2001-du}. The unique placement of a read in an assembly graph, hence, requires read length to be longer than the repeat length such that unique sequences not found elsewhere in the genome flank the repetitive sequence in the read. Gaps and collapsed regions of the genome, hence, often result from regions of the genome where the repeat length is longer than read length. There are, however, not many repeats except for segmental duplications\cite{Bailey2002-xn}, higher order repeats (HOR) in centromeres\cite{Willard1985-bo} and palindromic sequences in sex chromosomes that are longer than ONT and CLR reads \cite{Skaletsky2003-sr}.

A new generation of assembly algorithms based on de Brujin graph\cite{Lin2016-vl}, string graph\cite{Myers2005-ei, Chin2016-at} and OLC\cite{Koren2017-cq} were developed to leverage these long reads and enable end-to-end assembly of microbial genomes\cite{Bashir2012-cs, Chin2013-hp} and large mammalian genomes\cite{Chin2016-at, Koren2017-cq}. Complete hydatidiform mole (CHM) 1 BAC clones, for example, were selected for hierarchical shotgun sequencing to close existing gaps in the human reference genome \cite{Huddleston2014-rs}. At the time, contigs produced from these new assembly algorithms had unparalleled contiguity as measured by contig N50 \cite{}. In addition, misassembles can be corrected, and contigs can be ordered and oriented into scaffolds using optical genome maps from Bionano Genomics \cite{Pendleton2015-ue}. Chromosome-length scaffold construction, more importantly, has become routine through Hi-C scaffolding\cite{Dudchenko2017-kb} and the ability to visualise\cite{Robinson2018-os} and manually inspect Hi-C contact matrix for assembly curation\cite{Dudchenko2018-yl}. Trio-sequencing\cite{Koren2018-wg} and single-cell strand sequencing data\cite{Porubsky2021-ct} have also been used to also construct haplotype-resolved assemblies. These chromosome-length scaffold, most importantly, are often comparable or better than existing reference genomes in both contiguity and completeness \cite{Matthews2018-tv}.

Ultra-long read library preparation from ONT and CCS library preparation from PacBio were two additional breakthroughs that transformed how \textit{de novo} assembly is performed today. Ultra-long reads (>100kb) have been particularly useful for closing gaps\cite{Jain2018-zh} and for full-length sequencing of overlapping BAC clones for assembly of human chromosome Y centromere\cite{Jain2018-mg}. Human centromeres are enriched with AT-rich 171 bp tandem repeats called $\alpha$-satellite DNA. Centromeric $\alpha$-satellite DNA organises into HOR structures that are several megabases in length. Despite their crucial role in cell division, the organisation and structure of human centromeres were inaccessible to interrogation until the introduction of ultra-long reads. It is worth mentioning that centromere of b37 and hg38 reference genome exists as missing sequences and is not a true representation of the underlying sequence, respectively \cite{Miga2014-uv}.

CCS read length and accuracy have been leveraged to reduce computational complexity of all-to-all pairwise read alignments and shorten genome assembly time \cite{Chin\_undated-ye} and to distinguish recently diverged haplotypes and repeat copies such as segmental duplications \cite{Nurk2020-gu, Cheng2021-ij}. CCS reads are, routinely, used to produce haplotype-resolved chromosome-arm length contigs. It is worth mentioning that assembly algorithms often assumes that the sample in question has a haploid genome. This assumption results in haplotype collapsed assemblies where the assembled haplotype is not present in the population \cite{Schneider2017-yo}. The completion of telomere-to-telomere (T2T) CHM13 (T2T-CHM13) genome, including the short arms of five acrocentric chromosomes and centromeric satellite array, has been the culmination of years of effort to produce gapless and error-free assemblies \cite{Nurk2022-dv}. These advancements allow us construct high-quality reference genomes for fraction of what it used to cost to build the human reference genome. The number of new plant and animal assemblies has burgeoned thanks to these developments \cite{}.

**\subsubsection{Full-length transcript sequencing}**

In contrast, to short-read sequencing that requires \textit{de novo} assembly of RNA reads to acquire full-length transcripts, long-read sequencing can be used to obtain full-length transcript without assembly. Long-read sequencing has been used to successfully identify new isoforms in tissues and novel gene fusions in cancers \cite{}. Single-cell isoform-sequencing has also been used to find new isoform, to define the transcriptome atlas and to quantify the transcript in combination with single-cell RNA sequencing. In addition, these full-length transcripts has been successfully used for gene annotation of newly assembled genomes \cite{}.

**\subsubsection{Germline and somatic mutation detection}**

To date, ONT, CLR and CCS reads have been successfully used for germline SNP, small insertion and deletion\cite{} and structural variation detection \cite{}. The lower base accuracy and higher per base sequencing cost has limited the use of ONT and CLR reads for SNP and indel detection. The longer read length, however, enabled access to regions of the genome inaccessible with short reads and early success in identification of pathogenic mutations in undiagnosed patients with rare diseases \cite{}.

Structural variation detection with short reads relies on either changes in sequence coverage for copy number variation (CNV) detection and identification of discordant read pairs with aberrant distance and orientation for breakpoint, translocation and inversion detection \cite{Alkan2011-dv}. In contrast, long reads enable structural variation detection with nucleotide resolution through direct comparison of read and reference genome and is also more sensitive towards short tandem repeat (STR) expansions, short interspersed nuclear element (SINE) and long interspersed nuclear elements (LINE) insertion detection \cite{Chaisson2015-zz, Sedlazeck2018-oh, Denti2022-ux}. CHM1 CLR reads, for example, were also used to correct small misassembles in the reference genome and identify approximately 26,000 structural variations that were recalcitrant to detection using short reads \cite{Chaisson2015-zz}; the number of structural variations detected with long reads is at least double that detected with short reads. The number of structural variations is orders of magnitude smaller than the number of SNPs and indels, but structural variations alter greater number of bases and have a more pronounced impact on speciation and phenotype through gene regulation, duplication, translocation\cite{Weischenfeldt2013-tl} and conformational changes in three-dimensional genome configuration\cite{Spielmann2018-fm,}. In addition, complex structural rearrangements such as chromothripsis\cite{Stephens2011-gj, Korbel2013-to}, chromoplexy\cite{Baca2013-po} and templated insertions\cite{Yu2010-jr} are common oncogenic mechanisms. Repeat expansions and accompanied hypermethylation are common causes of neurological diseases\cite{Zhou2022-ci}. The severity of Parkinson’s disease, for example, is associated with repeat content and the size of the repeat expansion\cite{}. Single-molecule sequencing is the only reliable technology for repeat expansion detection. Low genetic diagnosis rate of approximately 30\% with short read sequencing and ability to detect haplotype phased genetic and epigenetic variations with single molecule sequencing has renewed interest to detect causal and putative pathogenic mutations in patients with rare genetic disease\cite{}.

Despite the advantages that long-read sequencing technologies offers compared to short-read sequencing technologies for somatic structural rearrangement detection, the application of long-read sequencing technologies to somatic mutation detection has been limited to date. There has been a handful publications that interrogated somatic structural rearrangements in breast cancer cell lines with long reads \cite{}. Somatic mutation detection with long reads is at the stage where we are re-creating the capabilities provided by short-sequencing technology and is not at the stage where we are finding somatic mutations that cannot be detected with short-read sequencing technology.