**\section{Single molecule sequencing}**

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 where event size is greater or equal 50bp. ONT and PacBio has fulfilled some of these promises, but Illumina still dominates the majority of the sequencing market as per base sequencing cost is still cheaper with the Illumina platform.

**\subsection{Nanopore sequencing}**

%% UCSC

%% ONT

%% blocked

%% limit

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

PacBio was founded in 2004 with aspirations to commercialize 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{}.

The metal film with the ZMW is placed on top of a glass and DNAP is immobilized at the bottom glass surface through surface chemistry modifications that prevents the adsorption of DNAP to the metal side walls \cite{}. 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{}. In addition, $\phi$29 DNAP was engineered through site-directed mutagenesis to use fluorophore-labeled deoxyribonucleoside triphosphate (dNTP) during DNA elongation \cite{}.

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 synthesized 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{}. 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{}. To date, DNAP kinetics has been used to detect including base modifications such as N6-methyladenine, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine \cite{} and DNA damage such as O6-mmethylguanine, 1-methyladenine, O4-methylthymien, 5- hydroxycyostine, 5hydroxyuracil, 5-hydroxymethyluyracil and thymine dimers \cite{}.

SMRT platform capability was initially limited to continuous long read (CLR) generation with 10-15% error rate instead of circular consensus sequence (CCS) generation with 0.1-1% error rate. 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. SMRT platform, hence, leveraged the improvements in DNAP processivity to first increase read length and subsequently, improve read accuracy.

Hybrid startegies

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 characterization 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 and structural variation detection.

Genomes are peppered with repetitive sequences. These repetitive sequences, for example, account for approximately 50\% of the human genome. 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. Collapsed regions of the genome, hence, often result from regions where the repeat length is longer than read length. There are, however, not many repeats except for segmental duplications, higher order repeats (HOR) in centromeres and palindromic sequences in sex chromosomes that are longer than ONT and CLR reads \cite{}.

A new generation of assembly algorithms based on de Brujin graph, string graph and OLC were developed to leverage these long reads and enable end-to-end assembly of microbial genomes and large mammalian genomes \cite{}. 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{}. Chromosome-length scaffold construction, more importantly, has become routine through Hi-C scaffolding and manual inspection of Hi-C contact matrix for assembly curation \cite{}. These chromosome-length scaffolds are often comparable or better than existing reference genomes in both contiguity and completeness \cite{}.

Reduce time for assembly, pairwise alignment

Ultra-long reads (>100kb) from ONT have also been particularly useful for closing gaps and for full-length sequencing of overlapping BAC clones for assembly of human chromosome Y centromere \cite{}.

centromeric alpha-satellite DNA

The centromere of b37 reference genome, for example, exists as missing sequences and centromere of hg38 reference genome is modeled on HuRef Sanger reads \cite{}.

The introduction of CCS reads

Human centromere sequence assembly

Human centromeres are enriched with AT-rich 171 bp tandem repeats known as alpha-satellite DNA

Human centromeres are mainly composed of **alpha satellite DNA** hierarchically organized as higher-order repeats (HORs).

Haplotype phased

Trio

CCS

The number of new plant and animal assemblies with chromosome-length scaffolds burgeoned thanks to these developments \cite{}.

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{}. CLR reads from CHM1 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{}; the number of structural variations detected with long reads is at least double that detected with short reads. Compared to short-read sequencing, long-read sequencing is also more sensitive towards short tandem repeat (STR) expansions, short interspersed nuclear element (SINE) and long interspersed nuclear elements (LINE) insertion detection \cite{}. The ability to detect structural variations

*In addition, the number of structural variations is orders of magnitude smaller than the number of SNPs and indels, but the cumulative sum of bases*

A new wave standardisation

The development of complementary scaffolding technologies and methods has allowed the rapid production of chromosome-length scaffolds [ref]. In addition, as the PacBio sequencing errors are thought to be introduced to the read randomly, the errors do not occur frequently in the same position and hence, highly accurate consensus sequence for the contigs could be achieved with an error rate between Q40-Q50.

These longer reads doubles the number of structural variations that can be detected from the comparison of a typical human genome to the reference genome and provides the identity of the structural variations at nucleotide resolution, in comparison with short reads that can only provide nucleotide resolution for small indels and that infers the size and position of the structural variation based on the presence of discordant reads, read features such as soft-clipped reads, and sequence coverage [ref]. The ability to determine the specific sequence at an insertion site is important as the sequence can determine the severity of disease [ref]. Moreover, instead of performing transcriptome-sequencing and de novo assembly of RNA reads to recover the transcript sequence, Isoform-sequencing allows for direct full-length sequencing of the transcriptome, enabling the discovery of new tissue-specific isoforms and isoforms is disease-specific [ref, ref]. The BAC clone assembly and tiling of gaps in the human reference genome. The use of ONT platform for full-length sequencing of BAC clones and assembly of human chromosome Y centromere.

The lower base accuracy and higher sequencing cost, however, limited the use of PacBio long-reads to de novo assembly and occasional discovery of pathogenic mutations in clinical sequencing settings. Together with the release of PacBio Sequel II platform, PacBio released SMRTcell with increased number of ZMWs from 1 million ZMWs to 8 million ZMWs and substantially increased the DNA polymerase processivity, which allowed for multiple sequencing of both the forward and reverse strand of the circulate template and construction of highly accurate consensus sequence, also known as circular consensus sequence (CCS) or high-fidelity (HiFi), reads. These CCS reads have an average read accuracy between Q20 and Q30 and have base quality scores ranging from Q1 to Q93. The higher base accuracy combined with longer read length enabled the construction of even more complete and contiguous assemblies with less sequence coverage [ref]. In addition, the higher base accuracy allows for de novo assembly algorithms to distinguish reads coming from two different haplotypes such that haplotype-resolved assemblies can be produced [ref, ref]. If de novo assembly algorithms are complemented with trio-sequencing, reads can be binned to paternal and maternal haplotypes to construct even more contiguous haplotype-resolved genomes and de novo assembly algorithms no longer must assume that the source of DNA is from a haploid genome. If trio-sequencing is not available, haplotagging can be alternatively performed to achieve haplotype-resolved assemblies.

%% Cornell

%% P0, P1, P2

%% Zero-mode wave guide

%% de novo assembly (OLC, string graph, haplotype phased)

%% T2T

%% pangenomes

%% scaffolding methods

%% haplotype phasing

%% deepvariant

%% imprinting

%% compound heterozygous

%% methylation

%% %% sequences in close proximity are in contact with each other more

%% conctact matrix can be used to discern correct assemblies from misassemblies

%% order and orient contigs %% matrix inversion, %% techniques from linear algebra

%% manually curate scaffolding and correct assemblies

%% pairwise sequencing

%% increase in compute power

In PacBio SMRT sequencing, DNA polymerase processivity and DNA damage repair during CCS library preparation, potentially is the most important factor for SMRT sequencing. DNA polymerase processivity determines the length of the template used for CCS library preparation and the base accuracy of the CCS reads. In addition, if DNA damage repair isn’t’ performed with sufficient fidelity, remaining DNA damage in the template can lead to earlier than expected termination of the SMRT sequencing process as DNA polymerase in the ZMW cannot perform DNA synthesis beyond the bulky DNA damage.

%% subread-throughput/single-strand read-throughput

**The Darwin Tree of Life (DToL) project**

These technological improvements in sequencing technologies, scaffolding technologies, cost of sequencing and de novo assembly algorithms have prompted many international collaborations to sequence and assemble and provide high-quality reference genomes of plant and animal genomes. The Wellcome Sanger Institute also initiated the 25 genomes project and the Darwin Tree of Life project to sequence and assemble all 66,000 eukaryotic species in Britain and Ireland [ref]. The DToL project has considered a number of sequencing technologies, but circular consensus sequencing has been adopted as the primary sequencing technology for de novo assembly.

Based on our understanding of duplex sequencing methods and the recently developed nanorate sequencing protocol, a derivative of the duplex sequencing protocol [ref, ref], and considering the similarities between two sequencing methods, we hypothesized that CCS reads might be as accurate or more accurate as duplex reads and can be used for single molecule somatic mutation calling. Both methods take advantage of the complementary information in the double-stranded DNA molecule to identify library errors and sequencing errors and to produce a highly accurate double-stranded consensus sequence (Fig 1). In the duplex sequencing protocol, a unique molecular identifier (UMI), consisting of 8 to 12 nucleotides, is added to the double-stranded DNA before the separation of the DNA molecule into single-strands for PCR amplification. Illumina adapters are added to the PCR amplified DNA, library is sequenced, and single strand reads belonging to the same DNA molecule is identified through their UMI. Because the forward and reverse strand is complementary and because DNA polymerase introduces errors infrequently during PCR amplification, library error and PCR amplification error should be present in one or two of the molecules, but not in all the molecules and most molecules should be consistent with each other such that the correct base can be ascertained. [bottleneck sequencing]. In contrast to the duplex sequencing protocol, CCS library preparation protocol doesn’t require a UMI as a single circular template is loaded to a zero-mode-waveguide (ZMW), DNA polymerase attaches to the template and sequences the circular template in a manner similar to rolling-circle-amplification to sequence both the forward and reverse strand multiple times until DNA polymerase terminates DNA synthesis [ref]. During DNA synthesis, DNA polymerase incorporates fluorescently labelled nucleotides and the fluorescence that is emitted during DNA synthesis is captured and fluorescence signal is converted to a nucleotide base. Thereafter, PacBio circular consensus sequence (pbccs) algorithm is used to generate CCS reads and for base quality score estimation.

If CCS reads have sufficient base quality score to enable single molecule somatic mutation calling and if DToL project generates high-quality reference genomes from these reads, we thought that we have an unparalleled opportunity to investigate the somatic mutational processes of non-human species at scale (discussed in Chapter 3). To enable, the study of somatic mutational processes in non-human species, we sequenced positive and negative control samples to assess the feasibility of single molecule somatic mutation detection with PacBio CCS reads, demonstrate that CCS reads indeed have sufficient base quality score for single molecule somatic mutation detection and develop and benchmark a method to call somatic mutations agnostic of clonality and species (discussed in Chapter 2). Based on our understanding of unique characteristics of CCS reads, we also sequenced and analysed granulocytes from Bloom syndrome patients and sperm samples of different ages to detect and analyse non-crossover and crossover resulting from mitotic and meiotic recombination at scale and across the whole genome (discussed in Chapter 4).

Here, we take advantage of CCS read length and high base accuracy to detect mutations at variant allele frequencies that was not possible before and to detect mutations resulting from double-strand break repair that could not be detected with short-read sequencing technologies.

%% \subsection{Somatic mutations}

%% \subsection{Mutational signatures and mutational processes}

%% darwin pondered the unit of inheritance (the physical material and the mechanism responsible for changing the physical material)

%% enodgenous and exogenous somatic mutation

%% DNA damage, repair, fixation

%% envrionment

%% DNA polymerase infidelity, germline mutations

%% importance of somatic mutation detection, lineage tracing, driver mutations

%% a harsh environment, insult to the DNA, necessary to repair DNA damage

%% type of DNA damage: single-base substitution,

%% what is universal about DNA? codon, degenerate, universal, 64 codons, stop-codon, start-codons

%% first protein

%% first riboenzyme?

%% first unicellular organism %% first in the sea

%% first lipid-bilayer that separates order from disorder, control of passage of molecules across a semi-permeable membrane

%% fusions, meiotic recombination, plant recombination?

%% non-hologous end joining

%% transcripion-coupled repair

%% Selection Pressure & Natural Selection & Survival of the fittest

%% deleterious, postivie,

%% linked by DNA

%% entropy to submission

%% Scientists still have not figured out how the first unicellar organism has arisen

%% Complexity that

%% DNA replication, DNA polymerase fidelity, DNA polymerase error rate, as a source of first mutations

%% first multicellular-organism

%% DNA nicks, DNA double-strand breaks, cyclo-butane dimer, UV light, chemicals

%% different types of DNA polymerases, redundancies

%% Oswald Avery: amino acids, greater number of combinations, genetic sequence as the transforming substance

%% Rosalind, Watson: Structure of DNA

%% what happened from the discovery of the structure of the DNA to the human genome project?

%% in humans

%% c-elegans? other species?

%% The Tree of Life is connected through genetic sequence

%% DNA is the puzzle that links us all

%% since inception, birth, somatic mutations starts to accumualte

%% fertilsiation for most organisms

%% cellcular division for unicellular organisms

%% depending on the timing and the type of tissue in which the somatic mutations occur somatic mutations are inherited to the daughter cells or the next generation

%% some mutations result in speciation

%% some mutations lead to survival of fittest

%% some mutations have a large consequence, recombination, structural variations

%% the study of mutations across the Tree of Life has been limited by the cost of reference genome construction and the availiabilty of reference genomes for population genetics and for comparative genomics.

%% the cost of reference genome construction has been prohibitively high

%% the human geome project, for example, cost 3 billion dollars, a dollar per base.

%% international collaboration, multiple sequencing centers with thousands of people

%% multiple-years

%% physical-maps %% fish %% restriction-enzyme based

%% YACs

%% fosmid 50kb-200kb

%% bacterial artificial chromosome clone 100kb fragments

%% gaps, missing sequences, acrocentric chromosomes, large sections of chromosome Y

%% unplaced, unlocalised chromsomes and contigs

%% placement of contigs, scaffolding of contigs

%% Sanger di-deoxy sequencing, limited to 500bp to 1000bp

%% Solexa and Illumina sequencing by synthesis

%% de brujin graph based assemblies are short, fragmented and incomplete

%% high-throughput, relatively high accuracy of short-reads

%% de novo assembly quality is a function of read depth, base accuracy, read length and complexity/repetitiveness of the target genome, %% Eric Lander

%% assemblies/genomes are abundant with sequences that are longer than Illumina read length: SINE, LINEs, repeat expansions, segmental duplications

%% longer read length is required to trasverse the repetitive sequence and uniquely locate/place the read amongst other reads, reads are collapsed into contigs in the face of high repetitive sequences

%% scaffolding technologies: mate-pair sequencing with longer-read inserts insufficient and not scalable

%% assembly and comparative genomics didn't improve in the last decade

%% cost was high, and the effort did not yield sufficiently meaningful assembly results

%% initially Single-molecule sequencing from Oxford Nanopore Technologies and Pacific Biosciences were also inaccurate and the read length were not magnitude of orders longer, low throughput

%% continuous long read sequencing from Pacific Biosciences, 10-15kb in read length with 10-15% error rate, the errors were thought to be random, free of amplification bias

%% sufficiently long enough to trasverse repeats, however not sufficient to distinguish between unique copies of segmental duplications

%% used to reconstruct missing sequences in the human reference genome %% eichler

%% updates in the human reference genome %% tina

%% CHM1 and CHM13 seuqencing to identify structural variations

%% pathogenic mutations/repeat expansions

%% ONT for chrY centromere sequencing

%% alpha-satelitte expansion

%% usefulness of haploid genomes

%% T2T consortium, for example, recently, completed the end-to-end assembly of CHM13 genome

%% high-throughput chromatin conformation capture sequencing (Hi-C), similar to mate-pair sequencing in concept, but across the whole-genome

%% 3C job-dekker, loops, configurations

%% originally used to study the three-dimensional genome configuration

%% chromosomes self-aggregate

%% end of one chromosome is in more contact with the end of the same chromosome than another chromosome

%% what about contacts between paternal and maternal haplotype of the same chromosome?

%% sequences in close proximity are in contact with each other more

%% conctact matrix can be used to discern correct assemblies from misassemblies

%% order and orient contigs %% matrix inversion, %% techniques from linear algebra

%% manually curate scaffolding and correct assemblies

%% studying the genomes from the Tree of Life provides snapshots of environments that the genomes were under through space and time

%% events that might have spurred natural selection, speciation and radiation

%% timed the emergence of species, but never timed the emergence of unique somatic mutational processes across time and space

%% assembly: assumption: haploid genome

%% Pacific Biosciences circular consensus sequencing, increase in the number of ZMWs per SMRTcell from 1 million to 8 million, circular consensus sequencing instead of continuous long read sequencing

%% increase in DNA polymerase processivity, continuous long-read sequencing perhaps once or twice per molecule, circular consensu sequencing: 8 to 16 times per molecule

%% because the errors are thought to be random, highly accurate circular consensus sequence generation is possible

%% estimated to have accuracy between Q20 and Q30

%% assemblies produced from PacBio CCS reads have accuracy between Q50 and Q60.

%% massive incerase in the contiguity and completeness and assembly of the genome

%% time to complete the genome

%% cost to complete the genome

%% thousands of scientists to handful of scientists

%% except for the most complex genome

%% significant upgrade in the quality of the genome compared to that produced from short reads

%% also comparable to that produced through the human genome project

%% or small organisms or unicellular organims with limited DNA %% low-input protocol makes this possible albeit with errors introduced during PCR amplification %% bias towards sampling of reads or amplification of dna molecules before library preparation

%% the number of eukaryotic species sequenced and assemblies with PacBio sequencing increased dramatically since the introduction of long-read sequencing

%% uncovering the evolutionary history of these species

%% Methods to study somatic mutations in cancer

%% the reasons to study cancer

%% somatic mtuational processes in cancer

%% mutational patterns, mutational signatures

%% tumour and matched normal

%% technical limitations of short reads

%% sub-cloncal

%% minute fraction

%% Methods to study somatic mutations in normal tissues

%% single-cell PCR amplification and sequencing

%% single-cell clone expansion and sequencing

%% duplex sequencing, nanorate sequencing

%% laser-capture and microdissection and sequencing of clonal tissues

%% driver mutations

%% drug resistance development

%% evolutionary history of cancers

%% developmental biology

%% lineage-tracing

%% Wellcome Trust Sanger Institute has initiated the Darwin Tree of Life project to sequence approximately ~66,000 eukaryotic species in the and the primary mode of sequencing is CCS sequencing, hi-c sequencing

%% to sequence and assemble the samples with CCS sequencing, scaffold the samples with Hi-C reads and to curate the scaffolded assemblies through manual inspection of the contact matrix

%% We and others have hypothesized the potential for CCS sequencing for somatic mutation detection

%% Nanorate sequencing, blunt-end restriction enzyme digestion, DNA nicks, dideoxy nucleic acid, DNA damage during sonication %% to preserve the native DNA molecule and to sequence the DNA molecule

%% We noticed the high simliarity between duplex sequencing and CCS sequencing and hypothesized that CCS sequencing might have sufficient base accuracy for single molecule somatic mutation detection, if we can distinguish highly accurate bases from that resulting from library errors, alignment errors and sequencing errors and systematic errors. artefacts that cannot be removed

%% Other mammalian species with shorter life span have higher somatic mutation rate such that at the terminal stages of life, the species in question have same mutation burden at the time of death

%% Peto's paradox

%% resequencing studies have enabled the identification of germline mutational process, somatic mutational process in humans

%% the study of other species have been limited to date

%% c-elegans? %% what are other species?

%% if our hypothesis is true, we conjectured that we will able to detect somatic mutations across the Tree of Life, reveal somatic mutational processes active in the species, time the emergence of somatic mutational processes and attribute the contribution of somatic mutational processes to the germline mutational process, %% environmental mutagenesis

%% in Chatper 2, we demonstrate that PacBio CCS base accuracy is sufficiently accurate to call and study single molecule somatic single-base-substitution across species

%% sequence samples with a single dominant somatic mutational process

%% know the mutational signature or have gold-standard mutational signature for the sample generated from single-cell clone expansion and sequencing

%%

%% somatic mutation detection from a single read alignment to the reference genome

%% if we were to call every mismatch between the read and reference genome, we will be able to call all somatic mutations at the cost of high false positive rate

%% Oxidative DNA damage

%% typically requires a normal sample to distinguish between germline and somatic mutations

%% typically requires multiple reads to suppport the somatic single base substitution

%% VCF file produced from somatic mutation callers are the sum of library errors, systematic errors, sequencing errors, alignment errors, %% reference bias?

%% unresolved errors

%% if we are able call somatic mutations from a single read alignment to the reference genome, we are not only able to reduce the cost of sequencing, but also do germline mutation calling from reduced read depth

%% 30X sequence coverage required to call heterozygous mutations %% reference

%% problems with PacBio CCS sequencing: incomplete removal of adapter sequences, chimeric sequences resulting from problems with adapater sequence calling, fragmer and concatmer

%% reads significantly shorter and longer than the read-of-insert length

%% empirically estimate the PacBio CCS base accuracy

%% PacBio CCS base accuracy has not been measured yet, PacBio CCS base also cannot be measured with exisiting sequencing technologies with lower base accuracy

%% in Chapter 3, confirm that our method is applicable to other eukaryotic species, we use the newly developed method to study somatic mutational processes across the ~400 eukaryotic species sequenced the Darwin Tree of Life project, attempt to understand both the germline and somatic mutational processes across species, identify potential sources of environemtnal mutagenesis

%% phorcus lineatus: age

%% insects: life cycle of insects (choleoptera)

%% mutation burden of insects with metamorphosis and without metamorphosis

%% germline mutational process

%% somatic mutational process

%% environmental mutagenesis

There are two types of third-generation sequencing: one from Pacific Biosciences and one from Oxford Nanopore Technologies and both companies attempt to sequence single molecule of DNA, in contrast to the sequencing by synthesis approach. These approaches, previsouly, had an error rate ranging from 20\% to 40\% depending on the library chemistry and the base caller version. PacBio introduced circular consensus sequencing in 2010, but circular consensus sequencing could not be adopted for mass-adoption as DNA polymerase for SMRT sequencing didn't have sufficient processivitiy to read long read-of-insert multiple times. Instead, PacBio offered continuous long read (CLR) sequencing to its customers which maximized for read length instead of average read accuracy. CLR reads typically have 10-15\% error rate, but is free from PCR amplification, the errors are thought to be randomly introduced and CLR reads have read length that is 100-fold longer than that from short reads. CLR reads, hence, was adopted for de novo assembly of complex genomes that could not be assembled with short reads and for structural variation detection. The longer read length enables the read alignment software to confidently assign the location of the reads relative to the reference genome as unique sequences are flanking repetitive sequences. Germline structural variation detection with long reads doubles the average number of structural variations discovered per genome compared to that from short reads and improves the diagnostic yield of rare genetic disease detection from 30\% to 80\%. The lower base accuracy and cost of SMRT sequencing, however, limited the wider adoption of PacBio SMRT sequencing except for one-off de novo assembly projects and clinical sequencing of patients with rare genetic diseases. PacBio, however, successfully engineered DNA polymerases with increased processivitiy and was further able to improve their circular consensus sequencing method such that read-of-insert with average read length of 10kb to 20kb can be read multiple times and because the errors are introduced randomly to each single-strand sequence templates, consensus sequence algorithms can take advantage of the complementary nature of double-stranded DNA to produce circular consensus sequences with average read accuracy greater than Q20.

%% in Chapter 4 and 5, we use the unique combination of long read length and base accuracy of PacBio CCS reads to study both meiotic and mitotic recombniation, respectively.

%% in Chapter 2 and Chapter 3, we demonstrate that PacBio CCS reads have sufficient read length and base accuracy for single molecule somatic single-base substitution agnostic of clonality and species.

%% to explore the unexplored phenomena of meiotic recombniation through Sperm PacBio CCS sequencing

%% diffences to previous attempts to understand meiotic recombination through trio sequencing and sperm-typing

%% gene conversions requires the detection of chimeric dna molecules with both maternal and paternal sequences

%% meiotic event generates 2 recombinant products and 2 wild type molecules

%% crossover leads to the generation of dna molecule with a stretch of paternal hetsnps followed by a stretch of maternal hetsnps and vice versa

%% gene conversion leads to the generation of a dna molecule where paternal hetsnps is flanked by maternal hetsnps (and vice versa)

%% complex recombinant product with resulting from both crossover and gene conversion

%% on average, there 1 SNP per 1000bp

%% requires long-range PCR products to detect

%% hotspots

%% coldspots

%% meiotic recombination product requires reads that can span multiple hetsnps and requires sufficient base accuracy to determine that hetsnp switch is a result of a biological event rather than a sequencing error.

%% in addition, meiotic recombination can be a source of mutagenic event

%% PacBio CCS reads have sufficient base accuracy to detect single molecule recombination events and associated mutations

%% recombniationi might not be a perfect/clean

%% mutational process that generates de novo single-base substitution seems to be driven by clock-like somatic mutational processes (SBS1 and SBS5)

%% mitotic gene conversion can be a source of oncogenic mechanism in somatic cells

%% simliar to meiotic recombination, products from mitotic recombination cannot be detected with short reads due to the technical limiations of the Illumina platform

%% sequenced Bloom syndrome patient samples with defects in DNA double-strand break damage repair process

%% known to have gene conersions or loss of heterozygous caused by gene conversions

%% perfect sample to assess the differences in mitotic and meiotic recombniation and gene conversions

%% mitotic gene conversions are thought to be longer in length

%% in Chapter 6,

%% the benefits of PacBio CCS sequencing

%% the last DNA sequencing platform

%% requires significantly less sequencing coverage than short reads to detect the same number of mutations

%% can detect small SNPs, indels, structural variations, 5mC from the same platform

%% with the development of himut, CCS reads can be also used to detect somatic mutations, gene conversion and crossovers from the same sample.

%% potentially other base modifications caused by environmental exposure, chemotherapeutics in the future

%% Moore's law: the number of transisitors per semiconductor has doubled, the distance at which the electrons has to be moved has shorteneed

%% the cost of sequencing per base was decreasing at a faster speed than Moore's law and many has anticipated that we might have a $100 genome, if the development had continued

%% stagnation in development, and Illumina monopoly status, financialisation, stock buybacks instead of research and develompent

%% increase in the number of ZMWs per SMRTcell, PacBio has achieved 8-fold improvement in throughput

%% increase in the read-of-insert length, doubling, stabiltiy of the circular template molecule

%% direct-engineering, directed-natural selection

%% increase in DNA polymerase processivity can increase either the read-of-insert length or the base accuracy of the same read-of-insert length

%% improvement in HMW DNA extraction, from the smallest organism

%% past Illumina platform generation has also required high DNA concentration

%% improvements in circular consensus sequence calling process can lead to the better discernment of adapter sequences from

%% PacBio CCS sequencing offers an alternative method for DNA sequencing with potential to improve throughput and base accuracy at a faster rate than that from Illumina unless Illumina profit margin compresses

%% PacBio CCS sequencing will be cheaper, more accurate, have higher throughput than Illumina sequencing

%% Illumina might compete in terms of price, but the wealth of information that is delivered from PacBio will be immense %% adoption curve

%% the cumulative improvement will us to better understand all of life

%% \subsection{The cost of Reference genomes as a bottleneck}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* %Second Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Reference genomes}

%% \subsection{De novo assembly}

%% bacterial artificial chromosomes

%% yeast artificial chromosomes

%% \subsection{Short-read sequencing}

%% de brujin raph

%% \subsection{Long-read sequencing}

%% linked-read sequencing

%% single-moelcule sequencing

%% oxford-nanpore technologies

%% Pacific Biosciences circular consensus sequencing

%% overlap-layout consensus

%% string graph

%% falcon

%% haplotype-phased

%% \subsection{Haplotype tagging}

%% \subsection{High-throughput chromatin conformation capture sequencing}

%% optical-mapping

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Third Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Resequencing}

%% \subsection{Germline mutation detection}

%% \subsection{Somatic mutation detection}

%% \subsection{Somatic mutation detection in cancer}

%% \subsection{Somatic mutation detection in normal tissues}

%% \subsubsection{Single-cell expansion and sequencing}

%% \subsubsection{Laser-capture microdissection and sequencing}

%% \subsubsection{Single-cell DNA PCR amplification and sequencing}

%% \subsubsection{Duplex sequencing}

%% \subsubsection{Pacific Biosciences circular consensus sequencing}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Fourth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Darwin Tree of Life project}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Fifth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Sperm sequencing for meiotic recombination product investigation}

%\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* % Sixth Section \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

%% \section{Bloom syndrome patient sample sequencing for mitotic gene conversion detection}

PacBio introduced circular consensus sequencing in 2010, but circular consensus sequencing could not be adopted for mass-adoption as DNA polymerase for SMRT sequencing didn't have sufficient processivitiy to read long read-of-insert multiple times. Instead, PacBio offered continuous long read (CLR) sequencing to its customers which maximized for read length instead of average read accuracy. CLR reads typically have 10-15\% error rate, but is free from PCR amplification, the errors are thought to be randomly introduced and CLR reads have read length that is 100-fold longer than that from short reads. CLR reads, hence, was adopted for de novo assembly of complex genomes that could not be assembled with short reads and for structural variation detection.

The longer read length enables the read alignment software to confidently assign the location of the reads relative to the reference genome as unique sequences are flanking repetitive sequences.

Germline structural variation detection with long reads doubles the average number of structural variations discovered per genome compared to that from short reads and improves the diagnostic yield of rare genetic disease detection from 30\% to 80\%. The lower base accuracy and cost of SMRT sequencing, however, limited the wider adoption of PacBio SMRT sequencing except for one-off de novo assembly projects and clinical sequencing of patients with rare genetic diseases. PacBio, however, successfully engineered DNA polymerases with increased processivitiy and was further able to improve their circular consensus sequencing method such that read-of-insert with average read length of 10kb to 20kb can be read multiple times and because the errors are introduced randomly to each single-strand sequence templates, consensus sequence algorithms can take advantage of the complementary nature of double-stranded DNA to produce circular consensus sequences with average read accuracy greater than Q20.