**\Chapter{Introduction}**

Somatic mutations can occur in cells at all stages of life and in all tissues. Somatic mutational processes in asexually reproducing cells and in the germline of sexually producing species are the source of genetic diversity upon which natural and sexual selection acts. The completion of the Human Genome Project (HGP) \cite{} and the advent of next-generation sequencing platform \cite{} have made it possible to detect and analyse somatic mutation landscape of thousands of tumours \cite{}. In stark contrast, the inability to detect somatic mutations in normal tissues at scale and the high cost of constructing high-quality reference genomes has prevented the characterisation of somatic mutational processes across the Tree of Life. The ability to produce long and accurate reads through circular consensus sequencing (CCS) from Pacific Biosciences (PacBio) has recently renewed interest for \textit{de novo} assembly of genomes for all of life \cite{}. Here, I describe a new method that uses CCS reads to enable the detection of somatic mutations present as a single copy from bulk normal tissue. I also discuss application of this method to discovering new somatic mutational processes across the Tree of Life.

**\section{The somatic mutation landscape of textit{H. sapiens}}**

To date, the study of somatic mutagenesis has largely been limited to \textit{H. Sapiens} and to somatic mutations present either in clones of cells as part of an ongoing effort to describe the driver mutation landscape in tumours. The contribution of somatic mutations to oncogenesis has been implicated as early as 1890 by Friedrich von Hansemann when he observed aberrant chromosomal alterations in cancer cells under the microscope \cite{}. The somatic mutation theory of aging also suggests that the acquisition of somatic mutations contributes to a decline in healthy cellular function and ultimately contributes to aging \cite{}. The continued decline in sequencing costs and the concurrent development of somatic mutation detection algorithms have enabled the detection somatic mutations in thousands of tumour samples \cite{} and subsequent functional studies have uncovered the precise mechanism with which these somatic mutations contributes to tumorigenesis. In more recent times, somatic mutation detection in normal cells has become increasingly important to lineage trace embryonic development \cite{} and to understand the transformation of normal cells to neoplastic cells \cite{}.

%% The advent of next-generation sequencing and the continued decline in sequencing costs have enabled us to sequence thousands of cancer genomes at scale and subsequent downstream sequence analysis has allowed us to discover tissue-specific driver mutations \cite{Martinez-Jimenez2020-kn}, identify biological processes that generate these mutations \cite{Alexandrov2013-kg}, to use somatic mutations as timestamps and biological barcodes to lineage trace development \cite{Behjati2014-gb}, to discover complex structural rearrangements such as chromothripsis \cite{Stephens2011-gj} that fundamentally changed the conventional view of tumorigenesis as the gradual process of the accumulation of somatic mutations \cite{Doll1954-of, Knudson1971-fg} and to better understand the relationship between abnormal embryonic development and paediatric tumour formation \cite{Marshall2014-ec}. International efforts such as the Cancer Genome Atlas (TCGA) program \cite{Weinstein2013-ko} and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium \cite{ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts} have also measured and analysed genetic, epigenetic, transcriptomic and proteomic aberrations in thousands of tumour genomes to understand how these aberrations contribute to the hallmarks of cancer \cite{Hanahan2000-dp, Hanahan2011-zr}.

**\subsection{Somatic mutational processes}**

Somatic mutational process is a continuous process throughout life, and multiple somatic mutational processes simultaneously act on the cell’s genome at any given time, leading to the generation and gradual accumulation of somatic mutations. The biochemical manifestation of a somatic mutation occurs in three distinct stages: 1) DNA damage or modification from either endogenous or exogenous sources, 2) mutation resulting from incorrect DNA damage repair and unrepaired DNA damage, and 3) the persistence of the mutation in the cell’s genome and its descendants \cite{Stratton2009-of}. Although most somatic mutations are considered benign (passenger mutations), some somatic mutations can confer a proliferative advantage to a cell and are classified as driver mutations. The detection of these driver mutations and their subsequent characterisation has been one of the primary motivations for cataloguing somatic mutations in cancer genomes. Somatic mutational processes discovered in \textit{H. sapiens} are further discussed in section .

**\subsection{Somatic mutation detection in tumours}**

Cancer is often described as the disease of the genome. The acquisition of driver mutations through a single event such as chromothripsis \cite{Stephens2011-gj} or gradual accumulation of somatic mutations \cite{Doll1954-of, Knudson1971-fg} is one of the primary contributors to tumorigenesis. Hence, somatic mutation detection is often the first step towards characterising the cancer genome.

Unlike germline mutations, where approximately 50\% and 100\% of reads will support the germline mutation, somatic mutations in tumour tissues can have a variant allele fraction (VAF) that ranges from 0\% to 100% based on their purity, cell clonality and changes in the number of chromosomes. As tumour samples often contain a mixture of both normal and cancer cells, the accurate measurement of somatic mutation VAF can present a challenge. In addition, the copy number of the chromosome can wildly fluctuate in the cancer genome as a result of chromothripsis \cite{}, aberrant chromosomal alternations \cite{} or loss of heterozygosity (LOH) \cite{}.

A matched tumour-normal sequencing is often performed to distinguish germline mutations from somatic mutations and to increase the confidence of each somatic mutation call. The presence of a matched normal also enables the calculation of tumour purity, the proportion of cancer cells in a tumour sample, which is another critical component that determines somatic mutation detection sensitivity. Germline mutations detected in the matched normal serves as a reference panel to determine whether the mutation detected in the tumour sample is a germline mutation or a somatic mutation. If a tumour sample, for example, has a low sequence coverage or low tumour purity, heterozygous germline mutation can be misclassified as a somatic mutation. Each somatic mutation detection algorithm uses a unique strategy to calculate the normal contamination in tumour \cite{}, tumour contamination in normal \cite{} and to differentiate somatic mutations from germline mutations. For instance, VarScan2 uses a hard filter \cite{}, MuTect uses a likelihood ratio \cite{Cibulskis2013-gw} and Strelka2 uses a mixture model \cite{} based on the number of reads supporting the mutation in the normal and tumour for classification of each somatic mutation candidate. Because each somatic mutation detection algorithm exhibits varying sensitivity and specificity and has different strengths and weaknesses, a consensus somatic mutation call from different somatic mutation algorithm is often used for downstream sequence analysis \cite{Bailey2020-ou}.

During library preparation, DNA damage is introduced to the template molecule and a set of DNA damage repair enzymes are used to repair the DNA damage. If DNA damage is left unrepaired or incorrectly repaired, the template DNA molecule is permanently altered prior to sequencing. During sequencing, base quality (BQ) scores are assigned to individual bases to the uncertainty of each base call. Somatic mutation detection algorithms depend on sequence coverage and BQ scores to calculate the genotype quality (GQ) score of a germline mutation \cite{} and the confidence with which the somatic mutation is called. Because sequencing instruments are agnostic to whether the base in the template molecule is derived from a somatic mutation or from a library error, BQ score does not necessarily reflect the probability that the non-reference base in read is the result of a mutation. Hence, low frequency library errors are often misclassified as somatic mutations \cite{}, and requires specialised filters to minimise the number of false positive mutations \cite{}.

PCR amplification, DNA oxidation and DNA crosslinking (in formalin-fixed tissues) commonly recognised sources of library errors \cite{}. The identification of oxidative DNA damage during DNA fragmentation through sonication, and the characterisation of read features that facilitate the differentiation of artefactual mutations from somatic mutations, serves as an illustrative example \cite{}. Acoustic shearing of DNA during DNA fragmentation oxidises guanine to 8-oxoguanine (8-oxoG) and the preferential pairing of 8-oxoG with adenine \cite{} is responsible for the generation of CCG>CAG/CGG>CTG transversions during PCR amplification \cite{}. The addition of DNA glycosylases during library preparation can ameliorate the effect of DNA oxidation. In addition, the sequence context and orientation bias of the mutation can also be assessed to determine the extent of the oxidative DNA damage in the library \cite{}.

The repeat content of the reference genome is another important factor for somatic mutation detection that is often ignored. Repetitive account for approximately 50\% of the human genome \cite{Lander2001-du}. If the repeat length is greater than the read length, read aligners cannot determine the location of the read with respect to the reference genome as the read could have originated from any copies of the repetitive sequence in the genome \cite{Li2008-dt}. Therefore, the accurate placement of reads requires repetitive sequences to be flanked with unique sequences not present elsewhere in the reference genome. Alignment errors in low sequence complexity regions is another common source of false positive mutations \cite{}. Consequently, both germline and somatic mutation detection algorithms discard mutations when the total number of mutations in close proximity to the called mutation, within a defined window, exceeds a predefined threshold \cite{}. As a result, the reference genome is divided into callable region and non-callable regions based on mappability of Illumina short reads and variant calling is often restricted to the callable regions of the genome \cite{1000\_Genomes\_Project\_Consortium2012-rj} and clinically relevant genes in non-callable regions are often excluded from analysis \cite{Wagner2022-ph}.

The completeness of the reference genome is another important consideration for somatic mutation detection. The human reference genome, for example, still remains incomplete with missing sequences, unplaced and unlocalised scaffolds and misassemblies such as erroneous sequence collapse and expansion \cite{Schneider2017-yo}. The telomere-to-telomere CHM13 (T2T-CHM13) genome, constructed from a combination of sequencing and mapping technologies, is currently the most accurate and complete human genome available to date. Compared to the human reference genome, the T2T-CHM13 genome has fully resolved sequences for the short arms of acrocentric chromosomes and centromeres of all chromosomes, with the exception of chromosome Y \cite{}. As expected, T2T-CHM13 improves the accuracy and precision of both read alignment and variant calling \cite{Aganezov2022-dv}.

To account for errors that cannot be eliminated through an analytical approach, a panel of normal (PoN) VCF file is generated from a set of normal samples, using the same parameters as those applied in somatic mutation detection. Afterwards, somatic mutation detections found in the PoN VCF is filtered out \cite{Cibulskis2013-gw}

**\subsection{Somatic mutation detection in normal cells}**

Somatic mutations are present in bulk tissue often as a single copy in a normal cell

The base accuracy and read length of Illumina reads, most importantly, is the common technical factor that limits the resolution at which the somatic mutations can be detected. MuTect, for example, cannot differentiate Illumina sequencing errors from low variant allele fraction (VAF) somatic mutations as a typical Illumina base call has a 0.01-1\% error rate \cite{Cibulskis2013-gw}.

Germline mutation rate

Somatic mutation rate

outs

Driver mutation

Passenger mutations

Mosaic mutations

**Somatic mutation theory of aging**

Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure \ref{}).

The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells.

To detect somatic mutations uniquely acquired in a single cell or present in few cells, several methods have been developed to increase either the base accuracy or the copy number of the mutant DNA above the limit of detection threshold.

Somatic mutation detection is challenging as most somatic mutations are present only once in bulk tissue samples, while most single-cell or single molecule technologies have too high an error rate to identify such unique events confidently.

Illumina’s technical limitations have limited somatic mutation detection to clonal or sub-clonal mutations. Two approaches have been developed to address these challenges: 1) to increase the copy number of the mutant DNA above the limit of detection threshold and 2) to increase the base accuracy of the Illumina reads through upstream changes in the library preparation protocol. Single-cell whole-genome amplification \cite{Lodato2018-hh}, single-cell clone expansion \cite{Lee-Six2018-qe} and laser-capture microdissection (LCM) \cite{Ellis2021-it} and sequencing adopts the former approach. Rolling circle amplification \cite{Lizardi1998-qh, Dahl2004-tm} and duplex sequencing methods \cite{Schmitt2012-yr, Abascal2021-pk, Hoang2016-jx} adopt the latter approach where a highly accurate consensus sequence is created from multiple copies of a single molecule.

Single-cell clone expansion and LCM sequencing are recognized as the gold-standard methods for somatic mutation detection in single-cells or clonal tissues, respectively. These methods have enabled the study of embryogenesis, somatic mutation rate, mutational processes, clonal structure, driver mutation landscape and earliest transformation of normal cells to neoplastic cells across a range of normal tissues, including adrenal gland, blood, bladder, bronchus, cardiac muscle, colon, endometrium, oesophagus, pancreas, placenta, prostate, skin, smooth muscle, testis, thyroid, ureter, visceral fat \cite{Lee-Six2018-qe, Martincorena2015-gu, Ju2017-vw, Martincorena2018-av, Brunner2019-xg, Lee-Six2019-vt, Yoshida2020-yr, Olafsson2020-vi, Moore2020-pi, Lawson2020-em, MSpencer\_Chapman2021-cq, Coorens2021-ct, Robinson2021-te, Grossmann2021-gd, Moore2021-dl, Park2021-fx, Ng2021-jd}. Duplex sequencing, however, is the most scalable option for ultra-rare somatic mutation detection and is the preferred method for circulating tumour DNA (ctDNA) based clinical applications \cite{Newman2016-cy}.

**\subsection{Somatic mutations as biological barcodes}**

to use somatic mutations as timestamps and biological barcodes to lineage trace development \cite{Behjati2014-gb},

Lineage tracing somatic mutation pediatric development

**\subsection{Somatic mutational processes}**

costs have enabled us to sequence thousands of cancer genomes at scale and subsequent downstream sequence analysis has allowed us to discover tissue-specific driver mutations \cite{Martinez-Jimenez2020-kn}, identify biological processes that generate these mutations \cite{Alexandrov2013-kg},

to discover complex structural rearrangements such as chromothripsis \cite{Stephens2011-gj} that fundamentally changed the conventional view of tumorigenesis as the gradual process of the accumulation of somatic mutations \cite{Doll1954-of, Knudson1971-fg} and to better understand the relationship between abnormal embryonic development and paediatric tumour formation \cite{Marshall2014-ec}. International efforts such as the Cancer Genome Atlas (TCGA) program \cite{Weinstein2013-ko} and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium \cite{ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts} have also measured and analysed genetic, epigenetic, transcriptomic and proteomic aberrations in thousands of tumour genomes to understand how these aberrations contribute to the hallmarks of cancer \cite{Hanahan2000-dp, Hanahan2011-zr}.

Multiple mutational processes simultaneously act on the genome at any given time and contribute to the accumulation of somatic mutations over an individual’s lifetime. To determine the mutational sources from a set of samples, mutational signature analysis is performed to either \textit{de novo} extract new mutational signatures or to assign the contribution of known mutational signatures to the mutation burden \cite{Alexandrov2013-fq}; a mutational signature is a mathematical abstraction of the likelihood that a particular biological process will produce a somatic mutation in a specific sequence context. During mutational signature analysis, somatic mutations are classified according to the event, the size of the event and the sequence context. Single base substitutions (SBS), for example, can be classified using the SBS96 classification system, which categorises SBS according to the six types of substitutions in the pyrimidine context (C>A, C>G, C>T, T>A, T>C and T>G) and the 16 possible trinucleotide sequence contexts derived from the 4 possible bases upstream and downstream of the substitution. SBS can be further subclassified based on their pentanucleotide sequence context (SBS1536 classification) and whether the SBS is located on the intergenic DNA, transcribed or untranscribed strand of the gene (SBS288 classification). Double base substitution, indel and structural variation classification system also exist for mutational signature analysis, but they are not the subject of interest in this chapter \cite{Alexandrov2013-fq, Li2020-vw, Steele2022-mn}.

**\subsection{Mutational signature analysis}**

The PCAWG consortium has discovered 67 single-base-substitution (SBS) mutational signatures \cite{Alexandrov2020-ys}. To date, the biological aetiology for 49 SBS mutational signatures has been determined (Table X). The discovery of new somatic mutational signatures is an ongoing process where the number and the aetiology of mutational signatures is constantly updated and refined with increase in the number of sequenced genomes. Genomics England and collaborators, for example, have leveraged 100,000 cancer genomes from around 85,000 patients to detect mutational signatures associated with rare and sporadic somatic mutagenesis \cite{Degasperi2022-qe}. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research \cite{Pich2019-ja, Aitken2020-sa}. Clinical sequencing of matched tumour and normal genomes is now routinely performed in the developed countries to help cancer patient treatment, fulfilling one of the many promises of the human genome project..

**\section{Single molecule sequencing}**

Single molecule sequencing method, unlike sequencing by synthesis method from Illumina, does not require the amplification of the template molecule and determines the nucleotide composition of an individual DNA molecule. Single-molecule real-time (SMRT) sequencing platform from Pacific Biosciences (PacBio) is one of the leading single molecule sequencing methods. Oxford Nanopore Technologies is an alternative provider of single molecule sequencing platform, but their platform is not the subject of this PhD dissertation \cite{}. During the early years of development, SMRT sequencing promised the following potential benefits:

\begin{enumerate}

\item lower input material for library preparation and sequencing

\item higher base accuracy

\item longer read length (10kb – 100kb)

\item production of contigs with higher N50

\item nucleotide-resolution identification of structural variations.

\item simultaneous detection of genetic variations and base modifications

\end{enumerate}

The first generation of PacBio continuous long reads (CLR) failed to meet these expectations. Illumina concurrently developed sequencing instruments that produced shorter reads with higher base accuracy, as well as cheaper per base sequencing cost \cite{}. Consequently, Illumina sequencing quickly emerged as the preferred method for sequencing. However, PacBio has started to fulfill some of these promises with recent developments that increases both read length and base accuracy.

**\subsection{Single-molecule real-time sequencing}**

PacBio was founded in 2004 with aspirations to commercialise 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, is at the heart of the SMRT platform. The ZMW acts as the sequencing unit and its unique properties helps 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 sequencing. 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 allow incorporation of 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 extension 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. An 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. Critically, 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 analysing 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 (IPD) \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}. 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. Circular consensus sequence (CCS) reads from the Sequel II instrument is reported to have 0.1-1\% error rate \cite{Wenger2019-pw} (Figure \ref{}).

\begin{}[h!]

\label{clr-ccs-sequencing}

\caption{}

\floatfoot{}

\end{figure}

The PacBio RS instrument with the first generation of polymerase and chemistry (P1-C1) produced CLR reads with an average read length of 1,500 bp with 10-15\% error rate \cite{} and the first generation of SMRTcells had 150, 000 ZMWs \cite{}. In contrast, the most recent PacBio Revio instrument generates CCS reads with an average read length of around 20,000 bp with 0.1-1\% error rate \cite{} and the latest SMRTcell has 25 million ZMWs. The sequence throughput of SMRTcells has increased exponentially from approximately 112 million CLR bases per SMRTcell to 250 billion CCS bases per SMRTcell, assuming that around half of the ZMWs are productive ZMWs. Thus, a single SMRTcell can generate 30-fold CCS sequence coverage of a human genome under one thousand dollars. The resulting output not only facilitates \textit{de novo} assembly \cite{} but also enables detection of haplotype phased \cite{} germline mutation \cite{} and base modifications \cite{}.

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

The high error rate of CLR reads and lower sequence throughput relative to Illumina sequencing instruments initially limited the applications of CLR reads to the assembly of microbial genomes \cite{}, targeted BAC clone sequencing to close gaps in the human reference genome \cite{Huddleston2014-rs}, and nucleotide-resolution structural variation detection for rare genetic disease diagnostics \cite{} or the study of hidden genetic variations \cite{}. The introduction of the P6-C4 chemistry and PacBio RSII instrument resulted in a substantial increase in CLR read length and sequence throughput \cite{}. This breakthrough has enabled the successfully assembly of larger and more repetitive mammalian genomes \cite{}, including those of great apes such as gorillas \cite{}, chimpanzees \cite{}, and orangutans \cite{}. CCS sequencing appears to be another advancement that allows us to explore the unexplored biological phenomenon.

**\subsubsection{Genome assembly}**

Genome assembly aims to determine to entire genetic information of an organism. Genome assembly can be divided into four distinct stages: 1) shotgun or hierarchical shotgun sequencing and quality control to remove reads from contamination, 2) all-to-all read alignments to find overlaps between reads and to connect overlapping reads into contigs 3) to use long-range information to order and orient contigs into scaffolds, 4) and to assess and finish the genome through gap closing.

The ability to determine the nucleotide composition of organisms at scale with ABI capillary sequencing platform initiated a race to determine the genome sequence of scientific and economic interest and to determine the method that is most suitable for the human genome project. In principle, genome assembly aims to use randomly selected DNA fragments from the genome, to find overlaps between the DNA fragments and to connect the overlaps into a single contiguous sequence. If the genome in question does not have repeats or if the read length is greater than repeat length, genome assembly becomes a trivial problem. Repeats account for less than X\% of prokaryotic genomes. Repeats, however, are common in eukaryotic genomes and account for ~50\% of the human genome. Repeats take many forms and repeats can exist as tandem repeats, palindromes, or inverted repeats. There are repeats created by retrotransposons where retrotransposons use copy and paste mechanisms to create copies of themselves in the genome. Segmental duplications is a special type of repeat where non-repetitive sequences greater than 1kb with interchromosmal or intrachromosomal duplications with sequence identity greater than 99% \cite{}. Simple repeats such as short-tandem repeat (STR) expansions where dinucleotides or trinucleotides exist as tandem repeats.

In addition, These repeats create false overlaps between reads and these false overlaps either leads to misassemblies such as collapsed haplotypes or to disconnected contigs\cite{}

Genomes are peppered with repetitive sequences. These repetitive sequences, for example, account for approximately 50\% of the human genome \cite{Lander2001-du}. Hence, the unique placement of a read in an assembly graph 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 often result from regions of the genome where the repeat length is longer than read length. Except for segmental duplications \cite{Bailey2002-xn}, higher order repeats in centromeres (HOR) \cite{Willard1985-bo} and palindromic sequences in sex chromosomes \cite{Skaletsky2003-sr}, long reads from the SMRT platform are able to span the most commonly occurring repeats such as $\sim$300 bp short interspersed nuclear element (SINE) and $\sim$5000 bp long interspersed nuclear element (LINE) \cite{}. Hence, long reads can be unambiguously aligned to the reference genome and can be used to resolve repeat-induced false overlaps in the assembly graph.

A new generation of assembly algorithms based on de Brujin graph \cite{Lin2016-vl}, string graph \cite{Myers2005-ei, Chin2016-at} and overlap layout consensus (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}. \textit{De novo} assembly algorithms perform all-against-all pairwise read alignments to identify overlaps between pairs of reads and the reliable overlaps are connected to produce contigs. The length of the overlap and the shared sequence identity between the overlap determines the reliability of the overlap. A repeat-induced overlap is generated when a read, derived from a repeat sequence, aligns to another read with a similar repeat sequence. If these repeat-induced false overlaps are left unresolved, fragmented contigs are generated from the assembly graph. As long reads often have unique sequences flanking the repeat sequence, repeat-induced overlaps that could not be resolved with short reads can be easily unentangled with long reads, enabling the generation of chromosome-arm level contigs \cite{}. CCS reads, with their higher base accuracy, excel at distinguishing more recently diverged repeats \cite{}. Long reads can also unravel divergences in the assembly graph, often referred to as a bubble, resulting from structural differences between the two haplotypes and produce haplotype phased contigs (haplotigs) \cite{}. As a result, the contigs produced from long reads have unparalleled completeness and contiguity compared to that produced from short reads \cite{}.

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 new assembly algorithms have been leveraged to complete reference genomes , to improve or replace exisiting assemblies

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}.

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}.

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 \cite{} 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, long 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 CLR reads for SNP and indel detection. However, the longer read length enabled access to regions of the genome previously inaccessible with short reads, and successful identification of pathogenic mutations in patients with rare diseases \cite{}.

Structural variation detection with short reads relies on 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 nucleotide-resolution structural variation detection through direct comparison of the read and the 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.

**\section{Tree of Life}**

**\subsection{Peto’s paradox}**

**\subsection{Darwin Tree of Life project}**

The advent of high-throughput long-read sequencing and genome mapping technologies \cite{}, improvements in base accuracy of long reads \cite{} and development of algorithms that take advantage of the longer read length and long-range genomic interactions \cite{} has brought new enthusiasm to sequence and assemble high-quality reference genomes \cite {}.

The Darwin Tree of Life (DToL) project is an ambitious project that aspires to construct chromosome-length scaffolds for 70, 000 eukaryotic species in Britain and Ireland \cite{}. In parallel, other international consortiums has initiated projects with similar aspirations for insects \cite{}, vertebrates \cite{}, invertebrates \cite{} and all of life \cite{}. The DToL project, currently, uses CCS reads for contig generation, Hi-C reads to order and orient contigs, and Hi-C contact matrix to manually inspect and correct chromosome-length scaffolds. We would like to highlight that the DToL project regularly updates their primary sequencing and mapping technologies and assembly, purging and scaffolding algorithms to reflect the advancements in the field. At the time of writing, the DToL project has sequenced approximately 800 species, completed the assemblies of approximately 500 species, and raw data and reference genomes have been made available to the public \cite{}.

**\subsection{Somatic mutation detection in non-human samples}**

**\section{Thesis objectives}**

To determine the germline and somatic mutational process across the Tree of Life, we considered the following:

\begin{enumerate}

\item Based on the similarities between the duplex \cite{} and CCS library sequencing \cite{}, we hypothesized that CCS reads might have sufficient base accuracy for ultra-rare somatic mutation and potentially single molecule somatic mutation detection.

\item CCS reads are reported to have a predicted accuracy above Q20, but their base accuracies have not been independently examined.

\item Somatic mutation detection algorithm needs to distinguish somatic mutations from germline mutations, in addition to, sequencing, alignment and systematic bioinformatic errors. We, unfortunately, cannot differentiate somatic mutations from library errors unless there are upstream modifications to the library preparation protocol.

\item Using samples with single ongoing somatic mutational process and mutational signature analysis, we can demonstrate that CCS reads have sufficient or insufficient base accuracy for single molecule somatic mutation detection and determine the parameters that influence sensitivity and specificity.

\item If the sample in question has either high mutation rate or high mutation burden, the expected and the correct mutational spectrum will be observable from the validation and test data sets, respectively.

\end{enumerate}

In short, we aimed to measure the CCS error rate, assess whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, develop a method to detect somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome and apply the method to understand germline and somatic mutational processes across the Tree of Life.