**\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{Weinstein2013-ko, ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts}. Although most somatic mutations are 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 to the hallmarks of cancer \cite{Hanahan2000-dp, Hanahan2011-zr} has been one of the primary motivations for cataloguing somatic mutations in cancer genomes. 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{}.

**\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 detect somatic mutations present in a clone of cancer cells. 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. Furthermore, 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{Nurk2022-dv}. 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}**

An individual begins to accumulate somatic mutations upon fertilisation and the gradual accumulation of driver mutations is one of the principal drivers of tumorigenesis \cite{}. To understand the mechanism underlying the transformation of normal cell to neoplastic cells requires the study of the driver mutation landscape in normal tissue across different ages and of individuals with genetic predisposition to cancer \cite{}. In addition, the ability to detect somatic mutations in normal cells is critical for earlier detection of cancer and monitoring of cancer relapse.

Somatic mutation detection in normal tissue presents a unique set of challenges as they are typically present as single copies in DNA extract from bulk normal tissue, except for mosaic mutations arising during embryonic development. Somatic mutation detection in tumour sample is possible as somatic mutation is present in a clone of cells above the 0.1-1% Illumina sequencing error rate \cite{Cibulskis2013-gw}. A newly acquired somatic mutation or somatic mutation with a VAF below the Illumina sequencing error rate is distinguishable from background noise. Hence, the detection of somatic mutation in normal tissue either requires an increase in the copy number of the mutant DNA above the limit of detection threshold or an increase in the base accuracy of the Illumina reads through upstream changes in the library preparation protocol.

**\subsubsection{Single-cell resolution somatic mutation detection}**

Both single-cell whole-genome amplification \cite{Lodato2018-hh} and single-cell clone expansion and sequencing \cite{Lee-Six2018-qe, MSpencer\_Chapman2021-cq} aims to increase the copy number of the mutant DNA prior to library preparation and sequencing. On the other hand, laser-capture microdissection (LCM) and sequencing leverage spatial partitioning of tissues through histological staining to isolate and sequence a clonal unit of cells, such as the colonic crypt \cite{Ellis2021-it}. Single-cell clone expansion and LCM sequencing are recognised as the gold-standard methods for somatic mutation detection in single cells or clonal tissues. In contrast, single-cell whole-genome amplification has several disadvantages, such as allele dropout, which occurs when one of the two alleles at heterozygous locus is not amplified, leading to a false homozygous mutation call. To date, single-cell clone expansion and LCM sequencing have been successfully used to determine the somatic mutation rate, the somatic mutational processes, and the clonal structure across a range 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}.

**\subsubsection{Single-molecule resolution somatic mutation detection}**

Unfortunately, single-cell clone expansion and LCM sequencing are arduous processes that are difficult to scale up for routine use in clinical settings. In contrast, liquid biopsies use duplex sequencing methods to detect driver mutations from circulating tumour DNA (ctDNA) in the plasma, enabling earlier detection of cancers and monitoring of tumour evolution \cite{Newman2016-cy}. Duplex sequencing methods generates multiple copies of the same template molecule and construct a highly accurate consensus sequence, leveraging the redundancies between the copies of the same template molecule and the complementary base pairing between the forward and reverse strand of a double-stranded DNA molecule \cite{Schmitt2012-yr, Hoang2016-jx, Abascal2021-pk}. The recently developed nanorate sequencing protocol enables single-molecule resolution somatic mutation detection \cite{Abascal2021-pk}. Duplex sequencing based somatic mutation detection is discussed further in chapter 2.

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

The fusion of the egg and sperm creates a zygote, whose genome has a unique combination of maternal and paternal alleles. The genome of the zygote orchestrates the programmed embryonic development from a single cell to a multicellular organism. Somatic mutations begin to accumulate with the first cell division of an embryo and are inherited by all the descendants of the embryonic stem cell that carries the mutation. Cells with the same somatic mutation are assumed to share the same stem cell lineage. Hence, somatic mutations have been used as biological barcodes to facilitate the lineage tracing of a cell and gain insight into embryonic development and the cellular origin of tissues \cite{Behjati2014-gb}.

The number of somatic mutations in a cell is a function of age and somatic mutation rate of the sample. Paediatric cancers from children, in contrast to tumours from adults, have a low mutation burden \cite{}. The hypothesis that tumours result from gradual accumulation of somatic mutations, therefore, does not apply to paediatric cancers. Abnormal embryonic development has been proposed as an alternative hypothesis to explain paediatric cancers \cite{}. In the former hypothesis, paediatric cancer will have unique set of somatic mutations, while in the latter hypothesis, paediatric cancer will share mosaic mutations and premalignant clones with the neighbouring normal tissue (Mosaic mutations are somatic mutations that occur during the first few cell divisions of embryonic development and shared across multiple tissues). The study of shared somatic mutations between bilateral and unilateral Wilms tumour and adjacent normal kidney tissues corroborates the latter hypothesis that paediatric cancer arises from aberrant foetal development \cite{}. A similar analysis in bilateral adrenal neuroblastoma revealed that the left and right adrenal gland tumours do not arise from a single premalignant clone, but rather two independent premalignant clones separated before the first few cell divisions of a zygote \cite{}. Intriguingly, these studies also show that the left-right axis of the kidneys and adrenal glands are established during the early cell divisions of a zygote.

*We estimate that approximately three base substitution mutations occur per cell per cell-doubling event in early human embryogenesis and these are mainly attributable to two known mutational signatures*[*7*](https://www.nature.com/articles/nature21703#ref-CR7)*. We used the mutations to reconstruct developmental lineages of adult cells and demonstrate that the two daughter cells of many early embryonic cell-doubling events contribute asymmetrically to adult blood at an approximately 2:1 ratio. This study therefore provides insights into the mutation rates, mutational processes and developmental outcomes of cell dynamics that operate during early human embryogenesis.*

*For example, somatic mutations arising in one of the two daughter cells of a fertilized egg will show VAFs of approximately 25% (*[*Fig. 1a*](https://www.nature.com/articles/nature21703#Fig1)*), compared to approximately 50% for inherited heterozygous polymorphisms, if the two cells have contributed equally to the adult tissue analysed*[*8*](https://www.nature.com/articles/nature21703#ref-CR8)*. To identify early embryonic base substitutions, we analysed whole-genome sequences of blood samples from 279 individuals with breast cancer (mean sequencing coverage 32-fold;*[*Supplementary Table 1*](https://www.nature.com/articles/nature21703#MOESM360)*), seeking mutations with VAFs ranging from 10% to 35%. To remove inherited heterozygous polymorphisms that fell by chance within this range, we phased candidate low VAF mutations to nearby germline heterozygous polymorphisms (*[*Fig. 1b*](https://www.nature.com/articles/nature21703#Fig1)*;*[*Supplementary Discussion 1*](https://www.nature.com/articles/nature21703#MOESM360)*). Substitutions present in regions with copy number variation were also excluded (*[*Extended Data Fig. 1*](https://www.nature.com/articles/nature21703#Fig5)*). After experimental validation by ultrahigh-depth targeted sequencing (median read-depth = 22,000;*[*Supplementary Table 2*](https://www.nature.com/articles/nature21703#MOESM360)*), we identified 605 somatic base substitutions with accurate VAF estimates (*[*Extended Data Fig. 2*](https://www.nature.com/articles/nature21703#Fig6)*) that appeared to be present in only a proportion of adult blood cells.*

**\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. 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}. As a result, each somatic mutational process leaves a characteristic imprint to the genome.

Mutational signature analysis is performed to determine the somatic mutational processes that have acted upon the genome and measure the contribution of each somatic mutational process to the mutational burden of the sample \cite{}. Mutational signature analysis can either \textit{de novo} extract new mutational signatures from a catalogue of somatic mutations from multiple samples or use a set of reference mutational signatures to calculate the attribution of each mutational signature to the mutation burden of the sample \cite{}. A mutational signature is a group of probabilities that represent the likelihood of a somatic mutational process generating a mutation at specific sequence contexts. The sum of these probabilities across all possible sequence contexts is equal to 1 \cite{}.

During mutational signature analysis, somatic mutations are classified according to the event, the size of the event and the sequence context. For example, single base substitutions (SBS) are often classified using the SBS96 classification system. In the pyrimidine sequence context, there are 6 types of substitutions (C>A, C>G, C>T, T>A, T>C and T>G) and 16 possible trinucleotide sequence contexts derived from the 4 bases upstream and downstream of the substitution. This combination of substitutions and trinucleotide sequence contexts generates the canonical SBS96 classification system. SBS96 classification can be further expanded to the SBS288 or the SBS1536 classification. SBS288 classification considers whether the somatic mutation has occurred on the intergenic DNA or transcribed or transcribed strand of the gene. SBS1536 classification adds additional 5’ and 3’ bases to the trinucleotide sequence context and further subclassifies SBS based on their pentanucleotide sequence context. Additional classification systems also exist for double base substitution (DBS), insertions and deletions (indels) and structural variations for mutational signature analysis, but they are not the subject of interest in this PhD thesis \cite{Li2020-vw, Steele2022-mn}.

The PCAWG consortium has discovered 67 SBS mutational signatures \cite{Alexandrov2020-ys} and the biological aetiology has been determined for 49 SBS mutational signatures (Table \ref{}, Appendix).

% \begin{table}

% \end{table}

The mutational signatures listed here are not an exhaustive list of all the discovered SBS mutational signatures. Ongoing research continues to find new somatic mutational signatures and determine the aetiology of these mutational signatures. For instance, Genomics England and collaborators 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}. Somatic mutations resulting from chemotherapeutic agents is another active area of research \cite{Pich2019-ja, Aitken2020-sa}.

**\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 the entire the 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) the use of long-range information to order and orient contigs into scaffolds, 4) and finishing the genome through polishing and gap closing.

The contiguity and completeness of the assembly is dependent on the repeat content and sequence coverage of the genome \cite{}. For instance, repetitive sequences account for approximately 50% of the human genome \cite{Lander2001-du}. If a read originates from a repetitive sequence, the read will overlap with another read with a similar or identical repeat sequence, creating a false overlap in the assembly graph. If a read longer than the repeat length is not present, repeat-induced overlap leads to either gaps or collapsed regions of the genome. To minimise the number of potential misassemblies, the human genome project (HGP) adopted hierarchical shotgun sequencing of 100-200kb bacterial artificial chromosome (BAC) clones to construct the human reference genome \cite{Lander2001-du}. The use of BAC clones reduces the simplifies the genome assembly process by reducing the complexity of the problem to a local assembly problem.

In contrast, shotgun sequencing based \text{de novo} assembly requires read length to be longer than the repeat length to construct high-quality reference genomes. Read length must be longer than the repeat length to ensure that unique sequences not found elsewhere in the genome flank the repetitive sequence in the read. This enables the unique placement of a read in an assembly graph. 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{}, reigniting interest for the development of \textit{de novo} assembly and scaffolding algorithms.

To complement the contigs produced from long reads, several new sequencing, physical mapping, and assembly and scaffolding algorithms have been developed. For instance, haplotype-resolved assemblies can be generated by leveraging the parent-specific kmers available through trio-sequencing \cite{Koren2018-wg} or through haplotype phasing using Strand-sequencing \cite{Porubsky2021-ct}. Genome maps from Bionano Genomics has been also used to correct misassemblies and to order and orient contigs into chromosome-arm level scaffolds \cite{Pendleton2015-ue}. Above all, chromosome-length scaffold construction has become routine through Hi-C scaffolding and the ability to visualise \cite{Robinson2018-os} and manually inspect Hi-C contact matrix for assembly curation\cite{Dudchenko2018-yl}. The ultra-long read (>100kb) library preparation and sequencing from ONT g \cite{} also facilitates the full-length sequencing of individual BAC clones \cite{Jain2018-mg} and closing of gaps in the genome that could not be assembled with CCS reads. The completion of the gapless T2T-CHM13 genome, including the short arms of five acrocentric chromosomes and centromeric satellite array, represents the latest advances in sequencing and assembly algorithms \cite{Nurk2022-dv}.

**\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 \cite{}, long-read sequencing can be used to obtain full-length transcript without assembly. Hence, long-read transcriptome sequencing is often referred to as isoform sequencing (Iso-seq). To date, Iso-seq has been used to successfully identify new isoforms in tissues \cite{}, novel gene fusions in cancers \cite{} and for gene annotation of new assembly genomes \cite{}.

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

To date, long reads have been successfully used for germline SNP and indel \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 long reads can access regions of the genome previously inaccessible with short reads and has significant advantage compared to short reads for structural variation detection.

Short-read based structural variation detection reads relies on changes in sequence coverage for copy number variation (CNV) detection \cite{} 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. Long-read based structural variation detection is particularly more sensitive towards short tandem repeat (STR), SINE and LINE insertion detection \cite{Chaisson2015-zz, Sedlazeck2018-oh, Denti2022-ux}. If the length and base accuracy of the long read is not sufficient to detect structural variations at the required resolution, \textit{de novo} assembled contigs facilitate the identification of larger and more complex structural rearrangements \cite{}.

CHM1 CLR reads, for example, were 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. Although the number of structural variations is orders of magnitude smaller than the number of SNPs and indels, structural variations alter greater number of bases and have a more pronounced impact on the phenotype of an individual \cite{Weischenfeldt2013-tl} and induces conformational changes in the three-dimensional genome configuration \cite{Spielmann2018-fm,}.

The diagnosis rate of rare genetic diseases is estimated to be approximately 30\% with short read sequencing \cite{}, which is not surprising considering the difficulty of their detection. In contrast, long read sequencing has repeatedly demonstrated its superiority for identification of pathogenic mutations. For instance, repeat expansions and accompanied hypermethylation are common causes of neurological diseases \cite{Zhou2022-ci}, and SMRT sequencing enables their simultaneous detection \cite{}. Moreover, STR expansion detection in patients with familial Parkinson’s disease with long reads has demonstrated that both the repeat sequence and the size of the repeat expansion is associated with the severity of the disease \cite{}. The ability to detect haplotype phased germline and epigenetic modifications has renewed interest to explore hidden genetic variation and to accelerate the identification of pathogenic mutations in patents with rare genetic diseases \cite{}. For instance, the human genome structural variation consortium is re-sequencing some of the samples from the 1000 genomes project with the SMRT platform to develop new structural variation detection algorithms and to study the genomic architecture \cite{}.

Although long-read sequencing technologies offer numerous advantages compared to short-read sequencing technologies for somatic structural rearrangement detection, long reads have been underutilised in somatic mutation detection. Only a handful of samples have been sequenced to interrogate somatic structural rearrangements using long reads \cite{}. Moreover, somatic substitution and indel detection algorithms, to our knowledge, have not been developed to leverage CLR or CCS reads. Hence, the author believe that a series of method development is required to enable the adoption of long-read sequencing for cancer genomics research.

**\section{Tree of Life}**

The Tree of Life is a recurring cultural and religious symbol that represents enlightenment, eternal life, and the universe. Charles Darwin first used the Tree of Life to illustrate the implications of natural selection and speciation (Figure \ref{}).

\begin{figure}

\caption{}

\floatfoot{}

\end{figure}

At the very root of the Tree of Life lies the last universal common ancestor (LUCA), from which all living organisms are believed to have descended. Each branching point depicts the emergence of a new group of species as a consequence of the struggle for existence. Each leaf, the endpoints of the branches, embodies distinct species and is the product of billions of years of evolution. Darwin used the Tree of Life as a symbolic metaphor to show the birth and death of new species and how different species are interconnected through a common ancestor.

Since the publication of \textit{On the Origin of Species} \cite{}, nucleic acid has been determined to be the carrier of genetic information \cite{}, the double helix structure of the DNA has been discovered \cite{}, and the same set of genetic code is used to code for amino acids across all forms of life \cite{}. These discoveries corroborate the conjecture by Darwin that all living species on Earth share a common ancestor.

The prohibitive cost of clone-by-clone sequencing for reference genome construction, the inability to produce high-quality reference genomes from short-read sequencing and the challenges of somatic mutation detection in normal tissues have prevented the detection and analysis of somatic mutations, the source of raw materials for natural selection, in other species.

**\subsection{Peto’s paradox}**

The study of somatic mutations and mutational processes in other species is another fascinating area of research to understand the underlying mechanism of aging and the incidence of cancer. Sir Richard Peto made the observation that there is no clear correlation between the number of cells and the incidence of cancer in other animals \cite{}. Because somatic mutational processes act on the genome of individual cells, and as the number of somatic mutations increases with age, larger animals with a greater number of cells and long-lived species are expected to have an increased risk of developing cancer. This, however, is not the case. Elephants are one such example. Elephants have more than one hundred times the number of cells pas humans, but they have a cancer mortality rate of 4.81%, compared to 11-25% cancer mortality rate in humans \cite{}. The analysis of the elephant genome has revealed that elephants possesses 20 copies of the TP53 gene, whereas the human genome contains 2 copies of the TP53 gene \cite{}, which is a tumour suppressor gene that is critical for maintaining genome stability \cite{}. The higher number of TP53 genes in the elephant genome is believed to be one of the primary reasons why elephants have a lower incidence of cancer.

**\subsection{Somatic mutation theory of ageing}**

The somatic mutation theory of ageing suggests that the gradual accumulation of somatic mutations leads to a decline in cellular function, which ultimately contributes to the ageing process \cite{}. The theory implies that shorter-lived species will have a higher somatic mutation rate than longer-lived species. Recently, the study of somatic mutations in colonic crypt across various mammalian species of different ages and lifespan have confirmed that the somatic mutation rate is inversely proportional to the lifespan of the mammalian species \cite{}. Hence, the longer-lived species have a lower somatic mutation rate. Intriguingly, the total mutational burden is similar at the end of their lifespan across different species. It is unclear whether this relationship between somatic mutation rate and the species’ lifespan holds true in other order, phyla, and kingdom.

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

The advancements in single-molecule sequencing technologies, concurrent development of new generations of \textit{de novo} assembly and scaffolding algorithms, and the ability to produce high-quality chromosome-length scaffolds at a fraction of the cost of hierarchical shotgun sequencing have brought new enthusiasm to assemble high-quality reference genomes for insects \cite{}, vertebrates \cite{} and invertebrates \cite{}.

The Darwin Tree of Life (DToL) project is an ambitious project that aspires to build reference genomes for all 70, 000 eukaryotic species in Britain and Ireland \cite{}. At the time of writing, the DToL project has sequenced approximately 800 species, completed the assemblies of approximately 600 species, and raw data and reference genomes have been made available to the public \cite{}.

I hypothesised that the availability of CCS reads and reference genomes through the DToL project presents a unique opportunity to discover new somatic mutational processes across the Tree of Life.

**\section{Thesis objectives}**

I have organised the remainder of this PhD thesis as follows. In chapter 2, I measure the CCS error rate and assess the potential for somatic mutation using CCS reads. I also introduce himut, a method that enables the detection of somatic mutations, agnostic of clonality and species. In chapter 3, I use himut to detect somatic mutations from eukaryotic species in the DToL project, describe the newly identified mutation signatures and measure the contribution of these mutational signatures to the germline mutational spectrum.