**\section{Conclusion}**

Here, I assess whether CCS reads are as accurate as duplex reads and demonstrate that a subset of CCS bases has sufficient base accuracy to enable single molecule somatic mutation detection using samples with single ongoing somatic mutational process. Himut takes as input a sorted BAM file with primary read alignments from bulk normal tissue, leverages CCS read length and base accuracy to distinguish somatic mutations from errors and germline mutations and returns a VCF file with somatic mutations. Mutational spectrum produced from aggregate of somatic mutations is concordant with the expected mutational signature from each positive control sample, showing that single molecule somatic mutation detection is indeed possible with CCS reads.

Using a cord blood sample with few somatic mutations, I examined the nature of residual false positive substitutions and associated CCS error profile that is shared across all samples. I empirically estimated that CCS Q93 base accuracy ranges from Q60 to Q90 depending on the substitution and trinucleotide sequence context, which is hundred thousand-fold to a billion-fold more accurate than Illumina bases and what enables somatic mutation detection with high confidence.

I conclude that false positive mutations are in fact derived from a combination of software errors. I show the persistence of inaccurate BQ score estimates using a modified pbccs that returns uncapped base quality scores, deepConsensus polished CCS reads and BQ score recalibration from partial order alignment between subreads and CCS reads from the same ZMW. I unexpectedly found that BQ score estimate becomes more inaccurate as the number of supporting subreads per CCS reads increases in contrast to the expected behaviour of the software (discussed and demonstrated in Chapter 3). In addition, I observe that false positive substitutions are enriched trinucleotide sequence contexts where the 5’ base or the 3’ base is identical to the substitution error. I hypothesize that inappropriate sequencing priors and underestimation of somatic mutations as potential sources of error in accurate BQ score estimation, and the use of trinucleotide sequence context HMM instead of dinucleotide sequence context HMM might ameliorate some of the issues. I, most importantly, show that subreads have sufficient base accuracy to generate CCS bases with $\sim$Q90 base accuracy at all trinucleotide sequence contexts, if there is enough supporting subreads per CCS read.

**\section{Discussion}**

I conjecture that issue with CCS BQ score estimation will be properly addressed and that majority of CCS bases will have $\sim$Q90 base accuracy in the imminent future. I, here, discuss the ramifications and potential applications following this development.

**\subsection{Somatic mutation detection}**

To date, CCS reads have been successfully used for construction of chromosome-length scaffolds of microbial and eukaryotic genomes \cite{}, used for germline SNP, indel and structural variation detection \cite{}, and have improved the genetic diagnosis rate of rare diseases \cite{}. The applications of CCS read for somatic mutation detection, however, have been limited and there has only been a handful of publications studying the complex structural rearrangements in cancers using CCS reads \cite{}. Here, I focused on single molecule somatic SBS detection with the intention to identify and analyse somatic mutational processes across the Tree of Life (discussed in Chapter 3) while others focused on improving the sensitivity and specificity of structural variations that could already be detected with Illumina reads \cite{}. Himut still cannot distinguish whether an individual SBS is an error or a somatic mutation, but posterior probability can be calculated to determine the probability that the substitution is derived from a biological process or a non-biological process (\ref{}).

This approach previously was used to determine SBS16 mutational signature, a signature associated with alcohol consumption, as the main source of somatic mutations in CTNNB1 gene in hepatocellular carcinoma \cite{}. Despite this problem, himut will still enable researchers to rapidly screen for mutational signatures from bulk normal tissue without arduous experiments such as LCM or single-cell clone expansion sequencing, identification of environmental mutagenesis such as exposure to aristolochic acids\cite{} across different locations and populations, lineage trace embryonic and tumour development through accurate detection of mosaic and somatic mutations, respectively. In addition, the ability to calculate the mutation burden in normal samples and thereby the age of the samples also raises the interesting question with regards to how to protect individual’s privacy when SMRT platform becomes the primary sequencing method.

Himut currently does not consider matched tumour-normal sequencing for somatic mutation detection, but this would be the natural next step as the number of matched tumour-normal samples sequenced with the SMRT platform is expected to increase with the introduction of the Revio instrument. In the future, when error-free native DNA CCS library preparation is possible and when CCS BQ scores are correctly calibrated, HMW DNA extraction, input requirements for CCS library preparation and sequence coverage of the sample becomes the limiting factor to identifying and studying somatic mutagenesis across all tissues and all species.

In the interim, I believe that a wider range of somatic mutation detection will be possible with the benchmarking approach I have established where a sample with a known double base substitution and indel somatic mutational process is sequenced and used to fine-tune the pbccs algorithm and improve himut sensitivity and specificity. UV light, for example, induces the photoexcitation and dimerisation of adjacent pyrimidines into cyclobutane pyrimidine dimer (CPD) and 6-4 photoproduct. Although the exact mechanism that converts DNA damage to DNA mutation is unknown, CPD deamination has been suggested as one of the mechanisms generating C>T mutations (SBS7abc) and CC>TT mutations (DBS1) \cite{}. Cisplatin, a commonly used chemotherapy drug, forms inter-strand DNA crosslinks to prevent DNA replication, which induces cell cycle arrest and apoptosis. Cisplatin produces a unique mutational signature where a single T insertion is introduced downstream of GG dinucleotides \cite{}, which is attributed to nucleotide excision repair of 1-3d(GpXpG) intra-strand cisplatin adducts \cite{}

**\subsection{Strand-specific somatic mutation detection}**

Somatic mutation is a three-step process: 1) DNA damage or modification from exogenous or endogenous sources, 2) failure to detect and repair the DNA damage correctly,and 3) fixation and persistence of DNA mutation in daughter cells. Mutational signature is a mathematical abstraction of these three inter-dependent processes (\ref{}) and describes the probability that a given somatic mutational processes will introduce a mutation at a specific sequence context.

The unique capability of SMRT platform to both generate single-strand consensus sequences (SSCS) and CCS reads, along with the ability to detect epigenetic modifications \cite{Vong2019-bi, Tse2021-or} and somatic mutations at a single-molecule resolution using CCS reads, presents an exciting opportunity to dissect each mutational signature in more detail.

DNA damage and repair process associated with SBS1 mutational signature, for example, is amenable to further qualitative and quantitative examination through CCS sequencing. The spontaneous deamination of 5mC to thymine results in a TG:GC mismatch, which results in C>T mutations at CG dinucleotides if left unrepaired by the mismatch repair (MMR) pathway. SSCS reads enable the detection of TG:GC mismatches and genome-wide mapping of DNA damage. CCS reads allow the simultaneous detection of 5mC base modification and C>T somatic mutations at single-molecule resolution (Figure \ref{}). If successful, we will be able to measure the \textit{in vivo} deamination rate and compare it against the \textit{in vitro} deamination rate of $5.8x10^{-13}$ per 5mC per second at 37°C \cite{}, measure DNA mismatch repair efficiency and fidelity under mutant and wild-type conditions, and estimate the nonlinear contribution of DNA damage, repair and mutation fixation process to the SBS1 mutational signature. MutS$\alpha$ deficiency, for example, elevates the number of C>T somatic mutations, demonstrating the critical role of MutS$\alpha$ in recognising the TG:GC mismatch and initiating MMR. Cross-examination of both SSCS and CCS reads and associated DNAP kinetics might also allow us to better understand the APOBEC mutagenesis (SBS2) resulting from cytosine to uracil deamination.

**\subsection{Gene conversion and crossover detection}**

I, here, also hypothesise that CCS read length and base accuracy can be leveraged for genome-wide gene conversion and crossover detection resulting from double-strand break (DSB) repair and characterise the differences between meiotic and mitotic recombination products.