**\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 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{}.

In the future, I believe that a wider range of somatic mutations will

The somatic mutations detected from our approach are not all true somatic mutations and if a user wishes to determine the confidence of the somatic mutation call or determine the posterior probability of the somatic mutation call, user can calculate the posterior probability of the substitution coming from a specific trinucleotide sequence context to have been generated by a specific and known mutational signatures [,ref, Eq].

In the future, when the CCS base quality scores are properly calibrated, single molecule somatic mutation detection might be truly possible.

PacBio has released new sequencing instrument Revio that increases the CCS read throughput 3 times with increase in read length and 3-fold increase in the number of ZMW, enabling the instrument to generate 30-fold sequence coverage genome at \$1000. This should drive adoption and increase the number of human genomes sequenced with the PacBio instrument. Researchers will typically use CCS reads for de novo assembly or for germline structural variation detection, but collection of CCS reads from public databases will enable the investigation of environmental mutagenesis across different populations across the globe and study the influence of germline mutation to somatic mutation generation and the combination of germline mutation and exogenous mutagen in generating new somatic mutagenesis.

**\subsection{Single-strand somatic mutation detection}**

Here, we did not focus on identifying and addressing the CCS library errors. We, however, believe that library errors must be present in CCS reads. HMW DNA shearing using XXX, for example, introduces oxidative DNA damage. 5’ filling or 3’ filling with XXX enzymes can perform strand displacement and use the template strand to synthesize the complementary strand, and these processes have been documented to generate library errors (ref, Nanoseq). To eliminate the library errors, HMW DNA could potentially be obtained from blunt-end restriction enzyme digestion, perform A-tailing and hairpin adapters could be ligated through blunt-end ligase. In addition, DNA molecules dependent on strand displacement and synthesis can be made not-viable for library preparation with the addition of dideoxy nucleotides or with DNA restriction enzymes that digests single-strand DNA.

PacBio CCS bases are at least hundred thousand-fold to one million-fold more accurate than Illumina short read bases.

Our method and CCS sequencing can be used to identify the presence of MMR for immunotherapy purposes.

In addition, the method is focused on somatic mutation detection from normal tissues but can be extended to matched tumour and normal settings to enable sensitive somatic mutation detection from tumour tissues. We also attempted somatic DBS detection, which occurs in ~100 fold less frequently than SBS, but like somatic SBS detection, true DBS signatures were outweighed by DBS artefact signatures.

We might be able to use a similar approach to also detect single molecule somatic structural variations.

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

We take advantage of the CCS base accuracy to detect gene conversions and crossovers in sperm samples and granulocytes from Bloom syndrome patients (discussed in Chapter 4). In addition,

HMW DNA input requirements for PacBio CCS reads limit the use of CCS sequencing for NIPT and ctDNA based genetic diagnosis (discussed in Chapter 5). HMW DNA input requirements are, however, expected to decrease with library preparation optimisation and like how DNA input requirements for Illumina sequencing has decreased.

During CCS sequencing, the kinetics of DNA polymerase during DNA synthesis is recorded. How fast, slow and whether the DNA polymerase paused during DNA synthesis is recorded. DNA polymerase kinetics data can be used to determine the base modification such as 5mC. Dennis Lo and colleagues, for example, have used ctDNA and NIPT DNA CCS reads to detect 5mC from single molecules and to successfully use them as diagnostic markers \cite{Vong2019-bi, Tse2021-or}. Single molecule somatic mutation and 5mC together should provide greater sensitive with which tumours are classified, monitor their evolution and their potential trajectory under selection pressure.

The introduction of himut allows researchers to detect 5mC, germline SNP, indel and structural variation detection and somatic mutation detection from a single SMRTcell on the Revio instrument. The breadth and depth of sequence and epigenetic information provided by CCS reads compared to Illumina sequencing for a single run of sequencing at a single molecule level should enable better diagnosis and study of samples.

Three Matrix = Mutational signature probability

Mutational signature is itself an abstraction of the three steps of somatic mutation: DNA damage, incorrect DNA repair and fixation. The accuracy of the PacBIo CCS bases and the ability to detect 5mC might enable us to dissect/deabstract the SBS1 mutational signature. The spontaneous deamination of 5mC to thymine (C>T) at CpG site is detected and repaired by the MMR repair machinery. We know the mutation probability of the spontaneous deamination of 5mC biological process to generate somatic mutations at CpG contexts, but we are, however, unaware of the rate at which spontaneous deamination of 5mC happens in vivo and the rate at which the C>T substitution is repaired and unrepaired by the mismatch repair (MMR) machinery. Using the base accuracy and the ability to detect 5mC base modification, we should be able to determine the rates of in vivo 5mC, success probability of the MMR machinery and the rate at which the C>T substitutions are fixed in the genome. We can imagine a scenario where a specific region will have wild type reads with 5mC, but one of the reads will have a C>T substitution. The subreads that was used to construct the CCS read can be examined to see whether the deamination happened on one of the strands and whether the other strand has complementary GC bases with 5mC. We can use similar approaches in the future to examine the probability of mutagen to generate DNA damage, DNA repair fidelity and DNA fixation probabilities.

The application for our method abounds as our method can act as a replacement for many of the laborious processes that provide single-cell resolution somatic mutation calls. Our method cannot provide single-cell resolution somatic mutation calls, but we can provide through time-series sequencing of the same sample, the monitoring of the same somatic mutation to study the population dynamics of the sample. In addition, our method can be used to screen for ongoing mutational processes in the sample cheaply without needed to perform laborious single-cell clone expansion and sequencing.

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