**\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, 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{}.

We examined mutational signatures in 324 WGS human-induced pluripotent stem cells exposed to 79 known or suspected environmental carcinogens.

1,2-DMH had a unique preponderance for C deletions flanked by T nucleotides.

Cisplatin produced a mutational signature characterized by T insertions at single T or long tracts of repetitive Ts. These T insertions were just downstream of GpG dinucleotides ([Figure 5](https://www.cell.com/cell/fulltext/S0092-8674(19)30263-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867419302636%3Fshowall%3Dtrue#gr5)C). This is in keeping with a previous report (

[Szikriszt et al., 2016](https://www.cell.com/cell/fulltext/S0092-8674(19)30263-6?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867419302636%3Fshowall%3Dtrue#bib50)) and is highly interesting because GpG dinucleotides are the targets of intrastrand crosslinks of platinum compounds.

Cisplatin treatment also induced a remarkable number of short insertion and deletion mutations, totalling 132 ± 34 per sample (Fig. [4a](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Fig4), Table [2](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Tab2), Additional file [1](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#MOESM1): Table S1). The insertions were almost exclusively one base long (95 % of all insertions, Fig. [4b](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Fig4)). We classified one-base insertions based on their sequence context (Fig. [4c](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Fig4), Additional file [1](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#MOESM1): Table S4). Ninety-four percent of one-base insertions were A/T base pairs. On the strand with the thymidine insertion, the preceding two bases were GG in 81 % of cases, presumably representing the site of an intrastrand crosslink Surprisingly, the bases following the insertion site also showed strong sequence preference. The first base following a thymidine insertion was 84 % T, while the first two bases together were 51 % TT. If the mutagenic process is DNA synthesis using the damaged strand as template, we can conclude that it preferentially inserts an extra adenosine when the bases 3’ to the template GG crosslink are thymines (see Fig. [4e](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Fig4) for a model). Six of the eight observed C/G base pair insertions occurred at CC/GG sites (Fig. [4c](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0963-7#Fig4)), also likely sites of intrastrand crosslinks.

In addition, 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.

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

T>G

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.

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.

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.

**\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,

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.

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).