**\section{Results}**

\**subsection{CCS library errors and sequencing errors}**

CCS reads have been successfully used for construction of highly contiguous and complete de novo assemblies \cite{} and germline mutation detection \cite{}. In these applications, the accuracy of individual base quality scores is not as important as ~50\% or ~100\% of the bases will support the consensus base, heterozygous or homozygous mutation. The accuracy of individual base quality scores, however, matters for ultra-rare somatic mutation detection as the base accuracy must be higher than the human genome somatic mutation rate (1-2 mutations per 1-4 weeks per cell), equivalent to approximately ~Q90 to distinguish sequencing errors from single molecule somatic mutations. In addition, library, sequencing and systematic errors and genomic DNA contamination are common sources of false positive somatic mutations.

We generated 30-fold CCS sequence coverage from BC-1, HT-115 and blood granulocytes from an 82-year-old female individual (PD48473b) and 70-fold CCS sequence coverage from cord blood granulocyte (PD47269d) with an average read length between 16 and 20kb (Table \ref{tab:CCS-sequence-statistics}) to achieve these objectives: 1), assess the potential for single molecule somatic mutation detection with CCS reads, 2) identify and address the sources of errors where possible and 3) empirically estimate the PacBio CCS error rate to define the limit of detection threshold, 4) develop a method for somatic mutation using CCS reads and 5) assess the sensitivity and specificity of our method.

\begin{table}[h]

\caption{Experimental Data}

\label{tab:CCS-sequence-statistics}

\begin{adjustbox}{max width=1.1\textwidth,center}

\begin{tabular}{l|cccc}

& BC-1 & HT-115 & PD47269d & PD4873b \\ \hline

Genomic DNA source & \multicolumn{2}{c}{Cell line} & \multicolumn{2}{c}{Blood granulocyte} \\ \hline

Age (years) & - & - & 0 & 82 \\ \hline

CCS read count & 5,962,252 & 5,933,281 & 12,156,251 & 4,949,180 \\ \hline

Mean length $\pm$ std (bp) & 18,571 $\pm$ & 17,038 $\pm$ & 16,523 $\pm$ 3,752 & 18,263 $\pm$ 1,753 \\ \hline

Q93 bases (\%) & 51.4 & 55.5 & 57.6 & 51.7 \\ \hline

Sequence coverage & 36.9 & 33.7 & 67.0 & 30.1 \\ \hline

Mutational process & APOBEC & POLE & \multicolumn{2}{c}{Clock-like} \\ \hline

Mutational signature & SBS2 & SBS10a, SBS10b and SBS28 & \multicolumn{2}{c}{SBS1 and SBS5} \\ \hline

Mutation burden per cell & $\sim$2,000 - 22,000 & $\sim$8,000 - 11,000 & $\sim$40 - 50 & $\sim$1400 - 1500 \\ \hline

\end{tabular}

\end{adjustbox}

%x\floatfoot{\small{CCS sequencing statistics, mutational process, associated mutational signatures and mutation burden are described for the negative control (PD47269d) and positive control (BC-1, HT-115 and PD48473b) samples.}}

\end{table}

We, first, examined the library preparation and circular consensus sequence construction process to minimise the number of library and sequencing errors. HMW DNA for CCS library preparation is often prepared through Qiagen Magattract or Circulomics HMW DNA extraction kit and HMW DNA is sheared to the appropriate size using a Megaruptor instrument. A hairpin adapter is attached to both ends of the double-stranded DNA molecule to create a topologically circular template. DNA nuclease is subsequently used to digest DNA molecules (e.g, failed ligation products) not suitable for sequencing. Primer with poly-A tail, thereafter, is annealed to the hairpin adapter sequence. BluePippin based size selection may additionally be performed to prepare size-selected libraries to maximize sequence throughput per SMRTcell.

A DNA damage repair enzyme cocktail (unpublished) is used to repair DNA damage (nicks, abasic sites, thymidine dimers, blocked 3’-ends, oxidised guanine and pyrimidines and deaminated cytosines) introduced during library preparation (personal communication). In addition, end-repair and A-tailing is performed to remove protruding ends and to enable adapter ligation, respectively. Defective DNA damage repair or unrepaired DNA damage manifest as library errors and can be misclassified as a somatic mutation. The precise identity of DNA damage repair enzymes in the cocktail are unknown. We, however, can make informed assumptions about their function and their impact on downstream sequence analysis, and highlight the DNA damage repair process that is most likely to introduce library errors. Nanoseq protocol, for example, pinpoints end-repair and nick translation process to be the primary sources of library errors. Strand-displacement synthesis during nick translation, for example, can introduce kilobases of sequences using the complementary strand as a template (Figure \ref{}) \cite{}.

CCS libraries are loaded on the SMRTcell and template DNA molecules diffuses into one of the ZMWs. A productive ZMW is defined as a ZMW with a single template molecule, from which a sufficient number of subreads are sequenced to construct a consensus sequence with at least Q20 average read accuracy. DNAP at the bottom of the ZMW binds to the DNA primer and initiates rolling circle amplification through strand-displacement synthesis. DNAP incorporates fluorescently labelled nucleotides, fluorescence emitted during DNA incorporation is measured and fluorophore is cleaved off upon successful incorporation. The wavelength of the fluorescence, length of the fluorescence, and duration between the successive pulses of fluorescence is used to determine the identity of the base and chemical modifications to the base.

The DNAP from the latest library protocol has sufficient processivity to generate an average of 10-12 full-length subreads on average for template molecule with read-of-insert length between 16kb and 20kb. The single-strand readouts of the forward and reverse strand of the template molecule are referred to as subreads. The first subread and the last subread are often partial readouts of the template molecule because of internal priming and sequencing termination, respectively while the subreads from the second to the second-to-last subreads are full-length readouts of the template molecule (Figure \ref{}). Assuming perfect detection of adapter sequences, odd-numbered subreads and even-numbered subreads are assumed to have the same sequence orientation as DNAP is agnostic to strand orientation. A draft sequence is constructed from multiple sequence alignment of subreads and is polished based on the realignment of subreads to the draft sequence. A dinucleotide sequence context Hidden Markov Model (HMM) is used to infer the base accuracy and DNA sequence from the observed subread bases (personal communication). A highly accurate consensus sequence can be constructed as sequencing errors are assumed to be randomly introduced without sequence context bias and are independent of one another. In addition, non-complementary base pairing between the forward and reverse strand indicates the presence of either a library or a sequencing error and resulting CCS is assigned a low BQ score.

PacBio circular consensus sequence algorithm (pbccs) calculates the median subread length and uses subreads with read length above 50\% of median subread length and below 200\% of median subread length for CCS generation for CCS generation. If adapter sequences are incorrectly detected within the subread or if adapter sequences are not detected where present, full-length subreads can be fragmented into multiple partial subreads and multiple subreads can be concatenated into a single subread, respectively. Unfortunately, read length based hard filters cannot identify all cases where adapter sequence detection has failed.

To identify potential errors introduced during CCS library preparation and sequencing, CCS and subreads from the same ZMW were analysed together and sequence quality control was performed (Methods). We observed that X\% of ZMWs have fragmented and/or concatenated subreads (Figure \{ref: }). We hypothesise that CCS reads with read length deviating from mean CCS read length are the result of failed adapter sequence detection and exclude these CCS reads from somatic mutation detection (Method). In addition, we also noticed higher than expected adenine and thymine proportion at the end of CCS reads resulting from incomplete adapter sequence trimming (Figure \{ref: }.

% CCS BQ score is biased towards Q93.

CCS reads have an average read accuracy of at least Q20 and individual BQ score ranges from Q1 to nominal Q93, corresponding to $0.5\times10^{-9}$error rate (Figure \{ref: }. To our knowledge, the accuracy of CCS BQ has not been examined to date. CCS read accuracy and BQ score is dependent on the number of subreads per CCS read (Figure \{ref: } and concordance between the subread bases and the CCS base. We confirm that the number of substitutions and indels is negatively correlated with CCS read accuracy and the number of subreads per CCS read as reported in a previous publication (Figure \{ref: }). The accuracy of the BQ score, hence, is expected to increase with the number of supporting subread bases. We, however, observed that the accuracy of the CCS BQ score decreases with increase in the number of subreads and that increase in the number of subreads per CCS read results in not diminishing returns, but negative returns to CCS base accuracy (discussed later in Chapter 3). To determine whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, we measured the empirical BQ score using cord blood CCS reads and (Methods) and ascertained that CCS bases have sufficient accuracy for rare somatic mutation detection where a sample has a high mutation burden or a high somatic mutation rate (Figure \{ref: }). Using positive control samples, we identified additional CCS read characteristics that influences somatic mutation detection sensitivity and specificity.

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

Here, we focused on single molecule somatic single-base substitution and the detection of larger structural variations that can only be detected with long-read sequencing is discussed in Chapter 4.

The somatic mutation spectrum of a normal tissue is continuous as somatic mutation accumulation starts post-fertilisation and as cells with driver mutations expand and colonise greater proportion of the tissue and somatic mutation is an ongoing process resulting from intracellular and extracellular sources (Figure XX). Hence, genomic DNA extracts from normal tissue is a combination of DNA molecules that has germline mutations and somatic mutations. To distinguish somatic mutations from germline mutations in a tumour sample, matched tumour and normal sequencing is performed, but we are attempting to separate the germline mutations from somatic mutations in a normal tissue.

To distinguish germline mutations from somatic mutations, himut traverses read across the chromosomes to first find candidate single base substitutions from a set of CCS reads that meets a set of pre-determined alignment properties and thereafter, determines whether the single base substitution is a homozygous reference allele, homozygous alternative allele, heterozygous allele, or heterozygous alternative allele (tri-allelic sites) using a Bayesian classifier identical to that MAQ and GATK uses for germline mutation likelihood calculation (Methods). Once the germline mutation status of the reference position is determined, himut only considers homozygous reference sites for SBS detection as other sites are candidates for somatic reversion and somatic reversions are not considered and somatic reversions might be the result of genomic DNA contamination. Himut, thereafter, applies a set of hard filters to mitigate the impact of the genomic DNA contamination and PacBio specific errors. To calculate the mutation burden of the sample, himut calculates the total number of trinucleotide sequence context that could have been potentially used for the somatic mutation calling with the same condition as somatic mutation calling and normalizes the mutation counts based on the trinucleotide sequence context frequency of the reference genome and callable bases (Methods). The user can prepare and supply a panel of normal VCF file to filter false positive somatic mutations resulting from systematic alignment errors and processing errors. In addition, true somatic mutations are haplotype consistent while false positive somatic mutations are haplotype inconsistent (Figure XX). To improve the sensitivity of sub-clonal somatic mutations, we take advantage of the CCS read length to haplotype phase the chromosome and use haplotype phased CCS reads for somatic mutation detection (Figure XX, Methods). Somatic mutation detection with short read sequencing uses adjacent hetSNPs to phase the somatic mutation and approximately ~30\% of somatic mutations are typically phased [ref, Serena’s breast cancer paper]. In contrast, the longer read length allows haplotype phasing 70\% of somatic mutations with CCS reads. In addition, to estimate the mutation burden of the sample, In the process of developing our method, we used the positive control samples to determine the features that are important for somatic mutation detection and suitable default parameters to be applied for future samples (Figure XX).

\subsection{Somatic mutation detection sensitivity and specificity}

Our method leverages the methods and approaches developed for germline and somatic mutation detection and improves upon them to apply our specific problem.

We applied our method to the positive control samples with different mutation burdens to obtain phased and unphased somatic mutations (Table 1). The mutation burden and mutational patterns from these samples were concordant to the mutation burden and signatures expected from these samples [Figure XX], demonstrating that PacBio CCS bases have sufficient base accuracy for single molecule somatic mutation detection.

Using mutational signature analysis, we were able to determine the specificity and sensitivity of our method. Using mutational signature analysis, we can determine the number of true positive somatic mutations that fits the expected mutational signature of the sample and what remains as the false positive somatic mutations; SBS2 signature is the only signature expected from the BC-1 sample and as a result, somatic mutations not attributable to SBS2 signature can be determined to be errors. Using the true negative, true positive, false negative and false positive somatic mutations, sensitivity, specificity and the F1 score of our method can be calculated. The number of true negative and false negative mutations can be determined from mutational signature analysis of filtered somatic mutations. We estimate himut to have XX\%, XX\% and X sensitivity, specificity, and F1-score, respectively. We, unfortunately, cannot compare himut with other existing somatic mutation callers as other callers are not designed for single molecule somatic mutation detection and/or somatic mutation detection is not technically possible.

The sensitivity improves from XX\% and XX\% and specificity increases from XX\% to XX\% when the grch38 human reference genome is used instead, reflecting that the higher quality assemblies leads to better variant calling.

In addition, we also assessed the impact of himut’s individual parameters to sensitivity and sensitivity independent of other parameters while other parameters are maintained as a constant. As expected BQ and germline GQ score has the greatest impact on himut sensitivity and other parameters have small, but positive impact on sensitivity and the incremental additive effects of all the parameters in the resulting specificity and sensitivity (Figure XX). Moreover, we also assessed the sensitivity and specificity of each parameter thresholds and generated receiver-operating curve for each parameter to determine the best default parameter for somatic mutation detection (Figure XX).

In the process, we found artefactual mutational patterns that occurs consistently across all samples, which we refer to as CCS artefactual signatures. To determine the sources of errors that produces the artefactual mutational pattern, we examined the CCS and subreads together. As the artefactual signature appears in all samples, we hypothesized those upstream systematic errors must be responsible for generating these sequencing errors.

\subsection{CCS error rate calculation and base quality score recalibration}

In contrast to the positive control sample, the cord blood sample should not have great number of somatic mutations and as a result, single-base substitutions detected from the negative control sample will be representative of the CCS error profile. The number of somatic mutations expected from the cord blood granulocytes are 40 – 50 somatic mutations per genome [reference Emily’s paper and other papers]. Our colleagues have also generated somatic mutations from single clone expansion and sequencing, the gold standard for single-cell somatic mutation detection and determined the ongoing mutational process in the cord blood granulocytes. The mutational pattern from cord blood granulocyte somatic mutations, unfortunately, was not concordant to what was expected from the sample, insinuating that the average CCS base accuracy is below Q93 as Q93 base should have been sufficient to capture all single molecule somatic mutations. We, however, used the false positive somatic mutations from cord blood granulocytes to determine the empirical CCS error rate. Using the cord blood HSC signature mutation probability and the trinucleotide sequence context count, we can estimate the number of somatic mutations expected from the sample, deduct that from the total called somatic mutations to calculate the number of mutations attributable to sequencing errors (Figure XX, Methods). We calculated the CCS base accuracy to range from Q60 to Q90 depending on the trinucleotide sequence context and the substitution (Figure XX, Methods)

We assumed that we have dealt sufficiently with the alignment errors and systematic errors in calling somatic mutation detection and wanted to determine the sources of errors upstream of germline and somatic mutation detection: library errors and sequencing errors. We did not focus on optimising the CCS library preparation to reduce the library errors as the Nanoseq protocol does to improve the duplex error rate. We, however, focused on identifying sources of sequencing errors. We hypothesized that CCS error rate must be resulting from incorrect CLR sequencing error priors. To test this hypothesis, partial order alignment between subread and CCS from the same ZMW was generated and we selected CCS bases with unanimous support from subread bases for somatic mutation calling (Methods). Somatic mutations called from CCS bases with unanimous support was concordant with what is expected across all the samples, suggesting that the inaccurate BQ score estimates are a software error and that this software error could be addressed with better subread substitution error priors. Google developed DeepConsensus to polish CCS reads with subreads and to re-calculate the BQ scores. DeepConsensus polished CCS reads have BQ score ranging from Q1 to Q50, and the estimates are too conservative compared our empirical estimations that can be derived (Figure XX). In addition, mutational pattern from Q50 somatic mutations is not concordant with what is expected from the sample, suggesting that the DeepConsensus polished CCS reads also don’t have accurate BQ score estimates.

In addition, the use of samples with single somatic mutational processes has the added benefit that these samples have been characterised in-depth through single-cell expansion and clone sequencing and we have determined the mutational probability of each substitution type in each trinucleotide sequence context. We, hence, are aware of the mutational pattern expected from the sample and can find the parameters that allows us to find mutational pattern from our positive control samples that is more consistent with what is expected from the sample. In addition, mutational signature analysis allows us to determine the number of mutations attributable to the correct biological process responsible for generating that somatic mutation and number of mutations attributable to false positive substitutions.

CCS BQ scores are capped at 93 as ASCII table doesn’t support higher BQ scores. We collaborated with PacBio to obtain pbccs that returns uncapped BQ scores and observed the uncapped BQ scores for problematic trinucleotide sequence contexts where false positive substitutions are abundant are still a problem, suggesting that the base quality score needs to be recalibrated.

The BQ score for CCS reads is capped at 93 as the ASCII standards cannot support higher scores and the user does not have access to the uncapped BQ scores. On average, DNA polymerase creates 10-16 subreads per CCS read per ZMW.

The number of subreads per CCS read is a function of DNA polymerase processivity, the rate at which DNA polymerase performs DNA replication and the read-of-insert length; The number of subreads per CCS read can either increase by increasing DNA polymerase processivity through protein engineering or by decreasing the read-of-insert length. The number of subreads and concordance between subread bases should be positively correlated with base accuracy. This, however, is not true in all circumstances and has unexpected negative ramifications as discussed in Chapter 3 and caution is required in choosing the read-of-insert length that will produce the CCS bases with the accurate BQ scores.