\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 mutation per 6 weeks per genome), 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}

Somatic mutagenesis is a continuous process throughout life. Bulk normal tissue has germline mutations that are inherited from parents, mosaic mutations that occurred during embryonic development and newly acquired somatic mutations from ongoing mutational processes. In addition, cells with driver mutations can outcompete neighbouring cells and undergo clonal expansion. Paired tumour-normal sequencing is often performed to distinguish germline mutations from somatic mutations in a tumour sample. Here, we present how we distinguish errors and germline mutations from somatic mutations in bulk normal tissue, leveraging CCS read length and base accuracy.

We, first, compared germline SNPs detected from both himut and deepvariant to assess whether our algorithm can accurately call genetic variations (Table). The number of SNPs and transition to transversion (TiTv) ratio is within the expected range, demonstrating that himut can also function as a standalone variant caller. We believe that algorithmic differences account for disparities in number of SNPs called with himut and deepvariant, which is a deep learning based variant caller that uses read pileup images for germline mutation detection while himut uses an analytical approach similar to GATK for germline mutation detection.

To distinguish germline mutations from somatic mutations, himut detects and classifies germline mutations as heterozygous, heterozygous alternative, homozygous alternative, or homozygous reference allele (Method, Figure \ref{}). Somatic mutation candidates are collected from CCS reads meeting the defined read-level thresholds and candidates are categorised according to their base-level conditions (Figure \ref{}). Somatic mutation detection is also restricted to homozygous reference allele sites as somatic reversions might be the result of DNA contamination. To calculate the mutation burden of the sample, himut identifies the number of callable bases using the same conditions as somatic mutation detection and normalises the somatic mutation count based on the number of callable CCS bases and reference bases (Method). A VCF file with haplotype phased heterozygous SNPs (hetSNPs), a VCF file with common SNPs and a PoN VCF file can also be optionally provided to call haplotype phased somatic mutations, to exclude false positive mutations resulting from DNA contamination and discard false positive mutations arising from systematic errors, respectively.

CCS read length and base accuracy can also be leveraged to phase hetSNPs and construct contiguous haplotype blocks, which enables haplotype phasing of CCS reads and haplotyped phased somatic mutation detection (Method). Read-backed phasing with Illumina reads uses adjacent hetSNPs to phase approximately $\sim$30\% of detected somatic mutations \cite{}. In contrast, haplotype phased somatic mutation detection with CCS reads uses all hetSNPs that CCS read spans and phases approximately $\sim$ 70\% of somatic mutations (Figure \ref{}). In addition, haplotype phased somatic mutation detection has three advantages: 1) CCS reads derived from DNA contamination often do not possess the same haplotype as the sample. If CCS read do not share the consensus haplotype, CCS read is excluded from somatic mutation detection (Figure, 2\ref{}) If two haplotypes are unevenly sampled, hetSNP can be misclassified as somatic mutations in low coverage samples. Restricting somatic mutation detection to haplotype phased regions limits somatic mutation detection to regions where both haplotypes have been adequately sampled. (Figure \ref{}), 3) CCS read with the same somatic mutations should share the haplotype and somatic mutation should not be present on both haplotypes (Figure \ref{}). Haplotype phased somatic mutation detection is especially helpful for samples with high heterozygosity.

\subsection{Somatic mutation detection sensitivity and specificity}

We called and benchmarked haplotype phased and unphased somatic mutations from the three positive controls with different mutational burdens and distinct mutational processes. Our unique benchmarking approach leverages the fact that a single somatic mutational process is active in each sample and that somatic mutation candidates are derived from either errors or newly acquired somatic mutations. We cannot be certain whether individual somatic mutations are derived from a biological process or a non-biological process, but mutational spectrum produced from the aggregate somatic mutations should be consistent with the expected mutational signature, if there is sufficient signal-to-noise ratio for somatic mutation detection. In addition, our approach is not biased towards Illumina callable regions of the genome unlike the Genome in a Bottle (GIAB) benchmarks \cite{} as our somatic mutation detection method is agnostic to reference position.

We calculated mutation burden from BC-1, HT-115 and PD48473b sample to be X, X, X, respectively, consistent with previous estimates \cite{}. In addition, high cosine similarity between the expected mutational signatures and mutational spectrum from our positive control samples demonstrate that PacBio CCS bases have sufficient base accuracy for rare somatic mutation detection where samples have a high mutation burden or high somatic mutation rate (Method). Moreover, we can determine the number of true positive mutations and false positive mutations from the called somatic mutations and the number of true negative mutations and false negative mutations from the filtered somatic mutations through mutational signature analysis. We can subsequently use these estimates to calculate the sensitivity, specificity, specificity, and F1-score for each of our sample (Method, Table). We also selected appropriate hard filter thresholds based on receiver operating characteristic (ROC) curve generated under a range of hard filter conditions (Figure \ref{}) and determined hard filters with the greatest impact on sensitivity based on odds ratio calculated in the absence and presence of the hard filter in question. BQ and GQ score were crucial for somatic mutation detection while other filters had marginally positive impact on somatic mutation sensitivity. We would like to also highlight that somatic mutation detection sensitivity and specificity increased when grch38 was used as a reference genome, reflecting better representation of genetic polymorphisms with improvements in assembly quality (Table). We, unfortunately, could not compare himut with other methods as himut is the first somatic SBS detection method with CCS reads and as somatic mutation detection below 0.1\% VAF has not been technically feasible with Illumina reads.

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

The mutation burden in cord blood sample is the lowest, with only 40-50 somatic mutations per cell \cite{}. CCS bases, unfortunately, does not have sufficient signal-to-noise ratio to enable somatic mutation detection in cord blood sample with high confidence. Mutational spectrum from the cord blood sample, which we refer to as the CCS error profile, is dissimilar to the expected mutational signature as the number of false positive mutations exceeds the number of true positive mutations (Figure \ref{}). CCS error profile occurs in multiple samples, suggesting that the error process is systematic in nature (Figure \ref{}). Using the number of false positive mutations and the callable number of bases, we calculated the CCS error rate to range from Q60 to Q90 depending on the substitution and the trinucleotide sequence context (Method, Figure \ref{}).

Library, sequencing, and software error upstream of somatic mutation detection are potential sources of false positive mutations. We triangulated software error as the origin of the CCS error profile through somatic mutation detection using uncapped BQ scores, deepConsensus polished CCS reads \cite{} and CCS reads with recalibrated BQ scores (Method, Figure \ref{}).

CCS BQ score ranges from Q1 to Q93 and BQ scores are encoded with the ASCII character encoding format. BQ score is capped at Q93 because ASCII character cannot support Phred-scaled quality values (QV) above 93. Inability to detect somatic mutations accurately with uncapped BQ scores demonstrates that there is a persistent problem with BQ score estimation (Figure \ref{}).

DeepConsensus calculates BQ score based on alignment of subreads to the CCS read from the same ZMW and BQ score of deepConsensus polished CCS reads ranges from Q1 to Q50 (Figure \ref{}), which we think is too conservative considering the empirical BQ score estimation from the cord blood sample. We also observed that somatic mutation detection with polished Q50 CCS bases did not generate the expected mutational spectrum while that with polished CCS bases with BQ score above Q30 generated the expected mutational spectrum, suggesting that once again BQ score is not accurately estimated.

To assess potential for single molecule somatic mutation detection with CCS reads, we performed partial order alignment between CCS read and subreads from the same ZMW and identifies bases where there is unanimous support for the CCS base from the subreads (Method). Somatic mutation detection with CCS bases with unanimous support from subreads generate mutational spectrum expected from the cord blood sample, suggesting that software error and not sequencing error is the source of false positive mutations. We hypothesise that consensus sequence construction and polishing algorithm consider somatic mutation as errors and as a result have incorrect sequencing error priors and BQ score estimates.