\section{Results}

\subsection{CCS read characterisation}

CCS reads have been successfully used for construction of highly contiguous and complete de novo assemblies and germline SNP, indel and structural variation detection for rare disease genetic diagnosis. 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 base accuracy must be higher than the human genome somatic mutation rate (1-2 mutations per 1-4 weeks per cell). Library, sequencing and systematic errors and genomic DNA contamination can be misclassified as somatic mutations, especially when a single read supports the alternative allele.

We generated 30-fold CCS sequence coverage from BC-1, HT-115 and blood granulocytes from an 82-year-old female individual (PD48473b) and 90-fold CCS sequence coverage from cord blood granulocyte (PD47269d) with an average read length of 15 to 20kb (Table 1, Figure XX) to achieve three objectives: 1), assess the potential for single molecule somatic mutation detection with PacBio 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.

To better understand the sources of sequencing errors, we first examined and identified sources of errors from CCS library preparation and sequencing. To create libraries with read-of-insert greater than 10kb, HMW DNA extraction is fundamental and is often carried out with either …, …, or Qiagen Magattract, or Circulomics HMW DNA extraction kit. If HMW DNA extraction is successful and if sufficient HMW DNA has been extracted, To create a topologically circulate template DNA, hairpin adapter is attached to the double-stranded DNA molecule (Figure X). DNA damage such as oxidative DNA damage introduced before or during library preparation is repaired using a cocktail of DNA repair enzymes (unpublished) and template DNA not suitable for sequencing is degraded using XXX DNase. The circular template, thereafter, is loaded to one of the ZMW in the SMRTcell and DNA polymerase at the bottom of the ZMW well initiates DNA synthesis using the circular template as a template. DNA polymerase incorporates fluorescently labelled free nucleotides, incorporation releases the fluorescent molecule, and the fluorescence is recorded through photonics and the wavelength of light emitted is recorded as one of the four nucleotide bases. DNA polymerase replicates the circular template through rolling circle amplification and sequencing terminates when DNA polymerase stops DNA synthesis. The DNA polymerase can initiate DNA synthesis from any starting points in the DNA template and equally terminate DNA synthesis from any point in the DNA template. Hence, the first and the last subread represents the partial readout of the template DNA while the second to the second subread are full pass subread that represents the full template DNA. DNA polymerase is agnostic to the strand orientation of the template DNA and as a result, odd-numbered subreads and even-numbered subreads are assumed to have the same sequence orientation. The draft consensus sequence is constructed from multiple sequence alignment of subreads, and the draft consensus sequence is polished through the realignment of subreads to the draft consensus sequence. Dinucleotide sequence context Hidden Markov Model (personal communication with PacBio staff scientists) is used to infer the underlying DNA sequence (hidden state) and the base accuracy from the observed subread bases [ref]. The concordance of the supporting subread bases with the consensus base determines the CCS base quality score.

To better understand the CCS construction, subreads and CCS reads from the same CCS reads were analyzed together. We noticed that XX\% of ZMWs have problems with adapter sequence detection, resulting in subread fragmentation and/or amalgamation (Figure XX); If the adapter sequence is incorrectly detected within the read-of-insert, the subreads can be split into multiple subreads and if the adapter sequence is not detected when present, two or more subreads can be connected to create a longer subread with both forward and reverse single-strand reads. CCS construction internally, hence, uses subreads that are longer than 50\% of the median subread length and shorter than 200\% of the median subread length. Despite this filter, full-length subreads are not purely selected and this filter doesn’t account for ZMWs where adapter sequences are incorrectly detected in all the subreads. This phenomenon might explain CCS read that deviate from the read-of-insert length and these CCS reads that deviate from the read-of-insert length might be error prone.

We performed additional quality control to understand CCS performance (Figure XX). The cumulative proportion of the nucleotide bases should be consistent across the length of the reads, but the higher proportion of adenine and thymine at the 5’ and 3’ end of the CCS read is the result of A-tailing and incomplete adapter trimming.

PacBio also reports that as the number of subreads per CCS read increases, the average read accuracy also increases. We also confirmed that the increase in number of subread per CCS read also increases the number of differences as measured by the number of substitutions and indels per CCS read (Figure XX). Moreover, as the number of subreads increase per CCS read, the proportion of Q93 base also increases, but unexpectedly the bases are skewed towards Q93 bases and as PacBio supports BQ score ranging from 1 to 93, CCS reads also not easy to compress. 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.

To date, CCS error profile has not been independently examined in depth

We initially used the positive control samples to assess whether Q93 CCS bases have sufficient base accuracy to enable single molecule somatic mutation detection and thereafter, used these samples to identify and assess features that influence sensitivity and specificity.

\subsection{Germline mutation and somatic mutation detection with PacBio CCS reads}

The sequencing statistics are summarised in Table 1. 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.