**Chapter 2: Single molecule somatic mutation detection with PacBio CCS reads**

**Introduction**

Somatic mutations can occur in cells at all stages of life and in all tissues. The biochemical manifestation of a somatic mutation requires three distinct stages: DNA damage or modification from either endogenous or exogenous sources, defective DNA damage repair and fixation, the persistence of the mutation in the genome of the cell and its descendants [ref]. Most somatic mutations are benign, but some confer a proliferative advantage and are referred to as driver mutations. Somatic mutation detection, hence, is often the first step towards characterising the cancer genome. The advent of next-generation sequencing and the continued decline in sequencing costs have enabled us to sequence genomes at scale and associated software development has allowed us to discover tissue-specific driver mutations, identify biological processes that generate these mutations, and to use somatic mutations as timestamps to lineage trace development [ref-ref]. Clinical sequencing of matched tumour and normal genomes is routinely performed in the developed countries to help patient treatment, fulfilling one of the many promises of the human genome project.

Somatic mutation detection, however, is not a solved problem. Somatic mutation callers, for example, employ different strategies and exhibit varying specificities and sensitivities. Consensus somatic mutation call, hence, is often used for downstream analysis [ref, Nature Communications]. The base accuracy and read length, of Illumina reads, most importantly, is the common technical factor that limit the resolution at which the somatic mutations can be detected. MuTect, for example, cannot differentiate Illumina sequencing errors from low variant allele fraction (VAF) somatic mutations as a typical Illumina base call has a 0.01-1% error rate [ref, MuTect2]. Library errors, introduced upstream of sequencing, is also often misclassified as somatic mutations [ref-ref, Science and NAR paper, Nanorate sequencing paper].

The repeat content of the genome is another hurdle for accurate somatic mutation detection. Repetitive sequences (e.g., tandem repeat expansions, retrotransposons, segmental duplications, telomeric repeats and centromeric alpha-satellite) account for approximately 50% of the human genome. If the repeat length is greater than the read length of the read with the repetitive sequence, read aligners cannot determine the reference genome location with high confidence as the read could have originated from any copies of the repetitive sequence. The accurate placement of reads, hence, requires repetitive sequences to be flanked with unique sequences not present elsewhere in the reference genome. Consequently, the reference genome is divided into callable region and non-callable regions based on mappability of Illumina short reads [ref, 1000G].

The completeness and contiguity of the reference genome is often ignored, but another important factor for somatic mutation detection. The human reference genome constructed from physical mapping, Sanger sequencing and scaffolding of bacterial artificial chromosome (BAC) clones with 50kb – 100kb is undoubtedly the best mammalian reference genome [ref, human genome project], but it is still incomplete.

The human reference genome, for example, still has missing sequences (also known as gaps), unplaced scaffolds, unlocalised scaffolds and mis-assemblies such as sequence collapse and expansion. Approximately 70% of the human reference genome is derived from genomic DNA of an anonymous individual of African-European ancestry [ref]. The current linear sequence of the human reference genome, therefore, may not accurately reflect the genomic diversity present in other populations and alternatively graph-based representation might better incorporate genomic diversity [ref, Ben, EKG, indel calling Rui]. The Genome Reference Consortium (GRC) has released grch38 build to address some of these issues. The Telomere-to-Telomere (T2T) consortium, alternatively, have generated gapless human assemblies using genomic DNA from complete hydatidiform mole (CHM) 13, long reads from Pacific Biosciences (PacBio) single molecule real-time (SMRT) platform and Oxford Nanopore Technologies (ONT) and high-throughput chromatin conformation capture (Hi-C) reads [ref, ref, ref]. T2T assemblies, as expected, improve the accuracy and precision of both read alignment and variant calling [ref].

Illumina’s technical specifications have limited somatic mutation detection to clonal or sub-clonal mutations, which in turn slowed our understanding of the transformation of normal cells to neoplastic cells and monitoring of tumour evolution and drug resistance development during cancer patient treatment. Two approaches have been developed to address these challenges: 1) to increase the copy number of the mutant DNA above the limit of detection threshold and 2) to increase the base accuracy of the Illumina reads through upstream changes in the library preparation protocol. Single-cell whole-genome amplification, single-cell clone expansion and laser-capture microdissection (LCM) and sequencing adopts the former approach [ref, ref, ref]. Rolling circle amplification and duplex sequencing (and its iterations) adopt the latter approach where a highly accurate consensus sequence is created from multiple copies of a single molecule [reviewed in ref, ref, ref, ref, ref]. Single-cell clone expansion and LCM sequencing are recognized as the gold-standard methods for somatic mutation detection in single-cells or clonal tissues, respectively. Duplex sequencing, however, is

Is the most efficient and scalable for option for ultra-rare somatic mutation detection and is the preferred method in most laboratories.

The duplex library preparation protocol starts with the sonication and fragmentation of genomic DNA and the attachment of 8 to 12 nucleotide unique molecular identifier (UMI) and Illumina adapters to double-stranded DNA molecules prior to their PCR amplification [ref]. The duplex library is often diluted before PCR amplification to achieve optimal sampling and duplication per template molecule [ref, ref BotSeq, Nanorate sequencing]. Illumina reads are subsequently grouped according to their UMI and are classified as Watson or Crick strand depending on whether the sequence was derived from Illumina adapter P5 or P7, respectively. A highly accurate double-strand consensus (duplex) sequence is constructed from the redundancies and complementarity between the forward and reverse strand reads; DNA polymerase, for example, might incorrectly replicate the template molecule, but the replication error will be present only in one copy or a subset of the copies. In addition, non-complementary base pairing between the forward and reverse strand will indicate the presence of replication errors. Consequently, duplex read promises theoretical base accuracy of 1 x 10-9 (Q90), but in practice achieves base accuracy of 1 x 10-6 (Q60) [ref, PNAS papers]

In contrast, duplex reads from the nanorate library protocol attains the promised Q90 base accuracy [ref]. To accomplish this, the nanorate library protocol identifies and addresses library errors upstream of PCR amplification to produce duplex libraries from error-free native DNA molecules; Genomic DNA, for example, is fragmented not through sonication, but using a blunt end restriction enzyme to prevent enzymatic DNA misincorporation during end repair and gap-filling. The addition of dideoxynucleotides also inhibits nick translation, rendering DNA molecules that require this process unsuitable for library creation.

PacBio CCS sequencing also take advantage of the redundant sequencing and complementary base pairing between the forward and reverse strand to construct highly accurate consensus sequences. The single-strand reads are referred to as subreads and an individual subread has 10-15% error rate. CCS reads are reported to have an average read accuracy between Q20 and Q30, but their individual base accuracies have not been examined to date. We and others have hypothesized that PacBio circular consensus sequence (CCS) reads might be as accurate or more accurate than conventional duplex reads based on the similarities between the two protocols [ref]. PacBio CCS base quality score ranges from Q1 to nominal Q93, representing error rate of 1 in 5 billion bases. If the base quality score estimates are correct, we imagined that genome-wide single molecule somatic mutation detection will be possible across all human normal tissues, agnostic of clonality as the human genome accumulates 1 to 2 somatic mutation per human genome per 1-4 weeks. If successful, haplotype phased germline mutation (SNPs, indels and structural variations), 5-methylcytosine (5mC) and somatic mutation detection will be possible from bulk normal tissue CCS sequencing. Our imagination inspired us to examine single molecule somatic mutations where a single read alignment supports the mismatch between the read and the reference genome. Our understanding of somatic mutational processes across different tissue types was critical in selecting the samples to assess and demonstrate the potential for single molecule somatic mutation detection with PacBio CCS reads.

International efforts such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium and normal tissue sequencing studies from independent labs have sequenced thousands of genomes and have identified hundreds to thousands of somatic mutations per genome [ref-ref]. Multiple mutational process simultaneously acts on the genome at any given time and contributes to the accumulation of somatic mutations over an individual’s lifetime. To determine the mutational sources from a set of samples, mutational signature analysis is performed to either de novo extract mutational signatures or to assign the contribution of known mutational signatures to the mutation burden; a mutational signature is a mathematical abstraction of the likelihood that a particular biological process will produce a somatic mutation in a specific sequence context. During mutational signature analysis, somatic mutations are classified according to the event, the size of the event and the sequence context. Single base substitutions (SBS), for example, can be classified using the SBS96 classification system, which categorises SBS according to the six types of substitutions in the pyrimidine context (C>A, C>G, C>T, T>A, T>C and T>G) and the 16 possible trinucleotide sequence contexts derived from the 4 possible bases upstream and downstream of the substitution. SBS can be further subclassified based on their pentanucleotide sequence context (SBS1536 classification) and whether the SBS is located on the intergenic DNA, transcribed or untranscribed strand of the gene (SBS288 classification).

The PCAWG consortium has discovered 67 single-base-substitution (SBS), 11 double-base substitution (DBS) and 17 indel mutational signatures, and has determined the biological aetiology for 49 SBS, 6 DBS and 9 indel mutational signatures [ref]. The SBS1 signature, for example, abstracts the spontaneous deamination of 5mC to thymine at CpG sites [ref]. The discovery of new somatic mutational signatures is an ongoing process where the number and the aetiology of mutational signatures is constantly updated and refined with increase in the number of experiments and samples studied. Genomics England and collaborators, for example, have leveraged 100, 000 genomes from around 85,000 patients to detect mutational signatures associated with rare and sporadic somatic mutagenesis [ref, serena’s paper]. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research [ref].

In this chapter, we assess the potential for single molecule somatic mutation detection using PacBio CCS reads, identify systematic errors with consensus sequence generation and base quality score estimation, propose potential solutions to address these issues. In addition, we detail the rationale behind the mechanics of himut and report its sensitivity and specificity. We have designed himut with ease of use in mind, and himut requires a sorted BAM file with primary read alignments and th as the only input and returns a VCF file with somatic mutations as output. We have released himut is available as a Python package under MIT open license at <https://github.com/sjin09/himut>.

We selected a set of samples (BC-1, HT-115 and granulocytes from an 82-year-old female individual) as positive controls and a sample (cord blood granulocyte) with little or no somatic mutations as a negative control to determine the artefact signature, empirically calculate the PacBio CCS error rate and the limit of detection threshold. In contrast to a typical sample where multiple mutational processes might be active at any given time, single-cell clone expansion and sequencing studies have definitively identified APOBEC, POLE, clock-like mutational processes to be the dominant ongoing somatic mutational processes in BC-1, HT-115 and granulocytes, respectively [ref, ref, Mia’s, Henry’s and Emily’s paper]. Single molecule somatic mutation candidates must either result from a biological process or from library, sequencing, alignment, or systematic bioinformatics errors. The concordance between the mutational pattern derived from the aggregate of somatic mutation candidates and the expected mutational signature can assess the specificity of the somatic mutation calls. If the mutational pattern, however, is discordant with the expected mutational signature, the sources of false positive mutations can be identified and addressed during the library preparation, consensus sequence generation and/or through downstream sequence analysis.

**Materials and Methods**

**Pacific Biosciences library preparation and circular consensus sequencing**

%% indel errors

%% homopolymers

%% continuous long read sequencing

%% error profile have not been characterised

Library preparation

Standard protocol

Pbccs

DNA primer

DNA polymerase

**Read alignment and germline mutation detection**

CCS reads were aligned to the human reference genome (b37 and grch38) with minimap2 (version --) with the parameters “” [ref] and primary alignments were compressed, merged, and sorted with samtools (version --)[ref]. Germline SNPs and indels were detected with deepvariant (version --).

**Germline mutation and somatic mutation detection**

Pysam [ref], cyvcf2 [ref]

We assume that

Parallelised

{ATGC}

Germline mutation detection

Somatic somatic mutation detection

Heteterozygous mutation

HetAlt

Homozygous alt

Homozygous reference

Doesn’t consider somatic reversions to the reference allele

Has the option to return haplotype phased somatic mutations

Contamination

**Panel of Normal construction**

**Haplotype phasing**

Haplotype phasing requires one to determine whether the polymorphisms are derived from a contiguous set of mutations. We treat haplotype phasing as a graph algorithms problem where each hetSNP is a node and measure haplotype consistency between a pair of hetSNPS to determine the validity of the edge. A single CCS read can span multiple heterozygous SNPs (hetSNPs) and a set of CCS reads can be used to measure the haplotype consistency between a pair of hetSNPs. Haplotype consistency if measured between all pairwise hetSNP and a pair of hetSNP is determined to be haplotype consistent through a binomial test (p<0.0001, one-sided). If a hetSNP is haplotype consistent with at least 20% of its possible pairs, hetSNP is a haplotype consistent hetSNP. Using the breadth-first-search algorithm, haplotype consistent hetSNPS are connected to construct a haplotype block and both haplotype consistent and haplotype inconsistent hetSNPs are returned as a VCF file.

**Haplotype phased somatic mutation detection**

**CCS read base quality score estimation and recalibration**

BAMsieve[ref]

Deepconsensus[ref]

CCS read and subread partial order alignment [ref] poatools

**Trinucleotide sequence context normalisation**

**Results**

**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.

Sequence identity

MAPQ

Alignment proportions

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.

**Germline mutation and somatic mutation detection method development**

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).

**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.

**CCS error rate 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 are 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.

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.

**Conclusion**

Here, we demonstrate that a subset of PacBio CCS has sufficient base accuracy to enable single molecule somatic SBS detection.

We estimate that CCS base accuracy ranges from Q60 to Q90 depending on the substitution and the trinucleotide sequence context. The CCS error rate is unexpectedly also dependent on the average number of supporting of subreads per CCS read (discussed in Chapter 3). The false positive substitutions resulting from inaccurate BQ scores are shared across samples and sequencing runs, suggesting that the issue is systematic in nature. Using a modified pbccs that returns uncapped BQ scores, we have confirmed that the same issue extends to CCS bases with BQ score above Q93. Google has developed deepConsensus to polish CCS bases and to revise CCS BQ scores based on multiple sequence alignments between subreads and CCS read from the same ZMW [ref]. deepConsensus BQ score estimates is capped at Q50, which is too conservative in comparison to our empirical calculation and similarly inaccurate as single molecule somatic mutation detection is not possible with deepConsensus Q50 CCS bases. We hypothesize the conservative deepConsensus BQ score estimate is since kmers arising from somatic mutations are treated as errors.

We observed that the false positive substitution is identical to the 5’ and 3’ and potentially the false positive substitution arises from the fact pbccs uses dinucleotide sequence context HMM and potentially a trinucleotide sequence context HMM might address the issue.

**Discussion**

To date CCS reads have ben successfully used for germline SNP, indel and structural variation detection and have improved the genetic diagnosis rate of previously undiagnosed rare diseases [ref, ref, Chaisson and Eichler, ngmlr, sniffles, deepvariant]. In addition, assemblies in combination with strand-seq enable detection of haplotype phased structural rearrangements longer than the read length [ref]. The applications of CCS read for somatic mutation detection, however, have been limited to date. Others have had limited success in using long reads for studying complex structural rearrangements in cancers and somatic retrotransposition detection [ref, ref]. The ability to detect large scale somatic structural rearrangements with long reads is especially important in determining the combination of genomic changes that results in the somatic structural variation. Here, we have focused on the successful detection of somatic SBS, but the method could be potentially improved to somatic indel detection. 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.

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.

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 [ref]. 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.

Darwin Tree of Life project has sequenced and assembled high quality reference genomes using CCS and Hi-C reads, providing us with the opportunity to detect somatic mutations from other non-human samples, for the first time (discussed in Chapter 3). The somatic mutation rate and mutational signatures are unknown across these species.

The study of somatic mutations across species allows us to tackel/attack the question posed by Peto's paradox. Why doesn't species with greater number of cells don't have higher incidence of cancer?

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,

Mutational signatures

\subsection{False positive substitutions}

\subsection{Environmental mutagenesis}

A cocktail of enzymes