**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

**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**

(Methods)

**Somatic mutation detection**

(Methods)

In addition, publicly available CCS reads from normal samples were processed to generate a panel of normal VCF file to filter false positives resulting from systematic errors.

**Haplotype phased somatic mutation detection**

**PacBio CCS error rate**

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.

**PacBio base quality score recalibration**

**Sensitivity and Specificity**

**Receiver-operating characteristics**

In addition, the use of samples with single somatic mutational processes has the added benefit that these samples has 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.

We need to define the problem first. In contrast to germline mutation detection where the mutation caller attempts to detect mutations that is homozygous or heterozygous, which exist as 100\% variant allele fraction and 50\% variant allele fraction, respectively, somatic mutations caller aims to detect somatic mutations that might be present in a single cell to a somatic mutations that might be present in all of the cancer cells and take into account tumour purity into the calculation. Somatic mutation callers, hence, often require a matched-tumour normal sequencing to distinguish germline mutations from somatic mutations and to calculate tumour purity??. In addition, because of techical limitations of short-read sequencing, low frequency somatic mutation with variant allele fraction below 0.1-1\% often cannot be detected.

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.

CCS read length and high base accuracy also allows us to haplotype phase the genome and call sub-clonal somatic mutations more effectively. If the detected mutation is a somatic mutation and not a sequencing error, the somatic mutation should be present on the same haplotype while that from sequencing error can be on both haplotypes.

To date, a number of approaches have been developed to detect somatic mutations using matched tumour-normal sequencing. Mutect..., Strelka2..., SomaticSniper..., Varscan2..., uses ... and Octopus..., to detect somatic mutations. To date, however, somatic mutation single-base substitution detection, have not been attempted with CCS reads and there has been limited effort in detecting somatic genomci rearrangements in cancer samples due to the availability of lack of scalable and reproducible methods.

%% BC-1, HT-115 cell lines

%% blood: SBS1, SBS5 + SBSX:

%% emily mitchell's paper

%% mia's paper

%% single dominant somatic mutational processes drive the generation of the most recently acquired somatic mutations

%% an individual/single substitution/difference between the read and reference genome cannot be determined to be a sequencing error or a true mutation that reflects the difference between the sample and the reference

%% if the circular consensus sequence base from Pacific Biosciences are sufficiently accurate, the differences between the read and the reference should be a reflection of biology and not the reflection of library errors, sequencing errors, systematic errors.

%% if single molecule somatic single-base substitutions, in aggregate, should be consistent with the expected mutational patterns from the sample if the mutations are correctly called.

%% sources of library errors, sonication, oxidative DNA damage

%%

%% single moelcule somatic single base substitution detection across all eukaryotic species, potentially sequenced with PacBio CCS sequencing method

%% even if soruces of errors introduced upstream of sequencing can be identified, it is hard to deal with them

%% systematic errors resulting from alignment errors and

%% Duplex sequencing

%% Nanorate-sequencing

%% Darwin Tree of Life project

%% Importance of somatic mutations: driver mutations, lineage trace development, time the emergence of driver mutations

%% somatic single base substitutions

%% small indels

%% structural variations > 50 bp

%% chromosomal translocations

%% tumour suppressor genes

%% technical limitations of illumina sequencing

%% normal sequencing: to understand the transformation of normal cells to neoplastic cells

%% DToL project

%% availability of reference genomes from diverse eukarytoic species

%% reference genome construction: CCS sequencing, linked reads, CLR, binano scaffolding

**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.

Inflated CCS base accuracy

Inflated CCS read accuracy

**Discussion**

\subsection{Liquid Biopsy}

\subsection{False positive substitutions}

\subsection{Environmental mutagenesis}

A cocktail of enzymes

has enabled us to detect more complex forms of genomic rearrangemetns such as chrmothripsis, chromoplexy, break-induced replication, fork stalling and template switching, etc and the elucidation of DNA damage and repair pathways that generate these complex rearrangements.

If single molecule somatic mutation detection is successful from PacBio CCS reads, we aimed to detect and characterise the somatic mutations and somatic mutational processes in 66,000 eukaryotic species from the Darwin Tree of Life project, which aims to sequence and assemble all the eukaryotic species in Britain and Ireland, providing an unparalleld opportunity to examine the somatic mutations and assocaited somatic mutational processes of non-human samples across the Tree of Life (disscussed in Chapter 3). 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?