Chapter 2

Single molecule somatic mutation detection

2.1 Introduction

Based on my understanding of duplex sequencing methods [39, 40] and the recently developed nanorate sequencing protocol [24], a derivative of the duplex sequencing protocol and considering the similarities between two sequencing methods, I hypothesised that CCS reads might be as accurate or more accurate than duplex reads and that they can be used for single molecule somatic mutation detection.

Most sequences have been derived by priming on both strands; this allows more confidence than when only one strand could be used [99]

[Frederick Sanger]

The Sanger sequencing method can be described as one of the first-generation of sequencing methods and the original duplex sequencing method. The first iteration of the Sanger sequencing method required a single-stranded DNA template, a primer designed to bind to the start of the template DNA molecule, DNA polymerase to bind to the primer and initiate DNA synthesis, and free deoxyribonucleotides (dNTP) and dideoxynucleotides (ddNTP) to elongate and terminate DNA synthesis, respectively. The chain-termination experiment is repeated multiple times with four dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP) to obtain DNA fragments of different sizes and DNA sequence is subsequently determined from reading the gel electrophoresis results from the four chain-termination experiments. Bi-directional Sanger sequencing can also be

performed to sequence both the forward and reverse strand of the template molecule and complementary base pairing between the two strands is leveraged to construct duplex reads with higher base accuracy [99]. To date, the Sanger sequencing method have been successfully used to obtain the 5,735 bp ΦX174 genome sequence [99] and reference genomes sequences of *D. melanogaster*, *C. elegans*, and *H. sapiens* [].

The current incarnation of the duplex sequencing method was developed for ultra-rare somatic mutation detection (<0.01% VAF) and to increase the limit of detection threshold beyond the technical limitations of the Illumina platform, in contrast to Sanger sequencing method that was used for genome assembly and germline mutation detection. The duplex library preparation protocol starts with the sonication and fragmentation of genomic DNA. Unique molecular identifier (UMI) consisting of 8 to 12 nucleotides and Illumina adapters are attached to double-stranded DNA molecules prior to their PCR amplification [39]. The duplex library is often diluted before PCR amplification to achieve optimal sampling and duplication per template molecule [40, 24]. PCR amplified library is sequenced using one of many Illumina sequencing instruments. 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 the P5 or P7 Illumina adapter, respectively.

A highly accurate duplex consensus sequence is, thereafter, generated leveraging the redundancies and complementary base pairing between the forward and reverse strand reads (Figure **??**). The higher sequence throughput of the modern Illumina instrument is critical in acquiring multiple reads (redundancies) from both strands of the template molecule and to identifying library errors introduced upstream of sequencing. DNAP, for example, might incorrectly replicate the template DNA molecule during PCR amplification, but the polymerase error will be present only in one copy or a subset of the copies. In addition, if both forward and reverse strands is sampled sufficiently, complementarity between the two strands can be used to demarcate bases with high accuracy from bases with low accuracy [39] and to estimate the base accuracy from the supporting bases and associated base quality scores [24]. Duplex reads, therefore, promises theoretical base accuracy of 1×10^{-9} (Q90), but in practice, duplex reads from the original protocol achieves a base accuracy of 1×10^{-6} (Q60) [39].

In contrast, duplex reads from the nanorate library protocol achieves the promised Q90 base accuracy and single-molecule resolution somatic mutation detection [24]. 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. Blunt end restriction enzyme, for example, is used to fragment gDNA to

prevent enzymatic DNA misincorporation during end-repair and gap-filling. In addition, dideoxynucleotides are added to prevent single-strand displacement synthesis through nick translation, rendering DNA molecules that require this process unsuitable for library creation (Figure **??**). A highly accurate duplex read, thereafter, is constructed as described above.

CCS sequencing like duplex sequencing also takes advantage of the redundant sequencing and complementary base pairing between the forward and reverse strand to construct a highly accurate circular consensus sequence. The single-strand reads are referred to as subreads and an individual subread typically has 10-15% error rate [100]. CCS reads are reported to have an average read accuracy above Q20 [64], but their individual base accuracies have not been examined to date. I and others have hypothesised 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 [] and the absence of PCR jackpot errors that occur in the earliest stage of PCR amplification. In addition, CCS reads have the added benefit of substantially longer read length (~10-20kb) that enables accurate placement of reads despite the presence of long repeats and allows more recently diverged repeats to be distinguished from each other in combination with the high base accuracy [].

CCS base quality score ranges from Q1 to nominal Q93, representing an error rate of 1 in 5 billion bases. If the BQ score estimates are correct, I imagined that single molecule somatic mutation detection will be possible across all human normal tissues, agnostic of clonality as the human genome accumulates ~17 somatic mutations per year per cell, equivalent to ~1 somatic mutation per human genome per 6 weeks [101]. In addition, in contrast to duplex sequencing methods where a matched normal sequencing is required to distinguish germline mutations from somatic mutations and where somatic mutation detection is limited to where restriction enzyme recognition site is available, CCS sequencing should enable genome-wide somatic mutation detection without a matched normal. If successful, haplotype-phased germline mutation (SNPs, indels and structural variations), epigenetic modifications and somatic mutation detection will be possible from bulk normal tissue CCS sequencing. This idea inspired us to assess the potential for single molecule somatic mutation detection using CCS reads 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 evaluate and demonstrate single molecule somatic mutation detection with CCS reads.

In short, I invert the premise that long reads are inaccurate and propose that CCS reads have the highest base accuracy among commercially available sequencing platforms. I assess the potential for single molecule somatic mutation detection using CCS reads, identify systematic errors with consensus sequence generation and base quality score estimation and propose potential solutions to address these issues. In addition, I present himut, a method that can call somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome. I detail the rationale behind the mechanics of himut and report its sensitivity and specificity. I have designed himut with ease of use in mind, and himut requires a sorted BAM file with primary read alignments as the only input and returns a VCF file with somatic mutations as output. Himut is publicly available at https://github.com/sjin09/himut as a Python package under the MIT open license.

Single molecule somatic mutation candidates are generated from either a biological process or from a non-biological process such as library, sequencing, alignment, or systematic bioinformatics errors. If a single read supports the mismatch between the sample and the reference, somatic mutation is indistinguishable from errors. If, however, there is sufficient signal-to-noise ratio somatic mutation detection, mutational spectrum produced from the aggregate of somatic mutations should be consistent with the expected mutational signature for the sample.

I selected a set of samples (the BC-1 and HT-115 cell lines, as well as normal granulocytes from an 82-year-old female individual) as positive controls and a sample (cord blood granulocyte) with few somatic mutations as a negative control to determine the limit of detection, empirically calculate the CCS error rate and describe the CCS error profile. 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 [102, 101]. The mutational spectra from previous studies and the contribution of different mutational signatures to the mutational spectrum serve as truth sets to unbiasedly assess the accuracy of our somatic mutation detection algorithm and to experiment and evaluate the impact of different hard filters to sensitivity and specificity.

The APOBEC family of proteins is part of the innate immune response to viruses and retrotransposon. APOBEC enzymes acts upon single-stranded DNA and RNA as cytidine deaminase and catalyses cytosine to uracil deamination to deteriorate and initiate the degradation of the viral genome []. APOBEC mutational process inadvertently introduces

C>T (SBS2) and C>G/C>A (SBS13) mutations to the genome at TCN trinucleotides (Figure **??**) [] and localised hypermutations called kataegis, which are often observed at chromothriptic breakpoints []. APOBEC mutagenesis is, in fact, observed in more than 50% of human cancers and accounts for considerable proportion of the total mutational burden [].

DNA polymerase α (POLA), δ (POLD) and ε (POLE) cooperate to perform DNA replication. POLA is responsible for initiating DNA synthesis while POLD and POLE is responsible for bulk of DNA synthesis with high fidelity on the lagging and leading strand, respectively []. POLD and POLE enzymes both have intrinsic proofreading capabilities and their 3'-5' exonuclease activity removes 3'-terminal misincorporated nucleotide. Replicative DNA polymerases still introduce errors every 10^4–10^5 nucleotides, but the mismatch repair (MMR) machinery corrects these errors. Individuals with inherited germline mutations or acquired somatic mutations that inactivate the POLE exonuclease activity have elevated somatic mutation rate and predisposes them to polymerase proofreading-associated polyposis, endometrial and colorectal cancers []. C>A mutations at TCN trinucleotides (SBS10a), C>A/C>T mutations at TCN trinucleotides (SBS10b) T>G mutations at NTT trinucleotides (SBS28) (Figure **??**) characterise POLE mutagenesis [].

Clock-like mutational processes are mutational processes that introduces mutations at a constant rate throughout life and hence, number of mutations attributable to clock-like mutational processes is proportional to the age of the individual. Clock-like mutational process is sample and species dependent, but C>T (SBS1) mutations at NCG trinucleotide (Figure **??**) and cell division independent background mutational process (SBS5) (Figure **??**) [] are determined to be clock-like mutational processes in normal human samples. C>T mutations at CpG dinucleotide result from spontaneous deamination of 5-methylcytosine to thymine and the unrepaired T:G mismatch manifests as somatic mutations. The exact aetiology of SBS5 is unknown, but somatic mutagenesis study in post-mitotic tissues such as neurons and smooth muscle suggests that SBS5 might be a cell division independent process and that SBS5 might be a manifestation of multiple different mutational processes []. .

2.2 Materials and Methods

2.2.1 CCS library preparation and sequencing

BC-1 and HT-115 cell lines were cultured in XX media containing XX and at XX in a humidified X environment. Umbilical blood (PD47269d) and peripheral blood sample of an 82-year-old female individual (PD48473b) were collected in 40-60mL lithium-heparin tubes and blood granulocytes were subsequently isolated using Lymphophorep. High molecular weight (HMW) DNA from BC-1 and HT-115 cell line and PD47269d and PD484873b blood granulocytes were extracted using Qiagen MagAttract HMW DNA extraction kit () and was sheared to 16-20kb DNA fragments using Megaruptor 3 system () with speed setting X. CCS sequencing libraries were constructed according to the 0.9.0 CCS library preparation protocol () and the libraries were sequenced using Sequel IIe instrument at the Wellcome Sanger Institute.

2.2.2 CCS read alignment and germline mutation detection

CCS reads with adapter sequences were identified with HiFiAdapterFilt [103] and were removed from downstream sequence analysis. CCS reads were aligned to the human reference genome (b37 and grch38) with minimap2 (version 2.24-r1155-dirty) with default parameters for CCS read alignment (-ax map-hifi –cs=short) [104] and primary alignments were selected, compressed, merged, and sorted with samtools (version 1.6) [105]. Germline SNPs and indels were detected with deepvariant (version 1.1.0) [106]. VCF files were compressed and indexed with tabix [107] and left aligned and normalised with bcftools (version 1.17-7-g097bda6) [108]

2.2.3 CCS empirical base quality calculation

To assess the potential for somatic mutation detection with CCS reads, we first assessed the accuracy of the BQ score estimate using CCS reads from cord blood granulocytes. The number of somatic mutations in cord blood granulocytes is limited to 40-50 somatic mutations per cell [109], and hence most SBS, excluding germline mutations, in cord blood granulocyte sample results from library, sequencing, alignment or bioinformatics error. The number of matches and mismatches were counted for each BQ score estimate to calculate the empirical BQ score. I considered reference allele and germline SNPs as matches and all other SBS as mismatches. Germline mutation detection using himut is described below. I excluded germline SNPs with genotype quality (GQ) score below

minimum GQ score of 20 and read depth above maximum depth threshold $4d+\sqrt{d}$, where d is the average read depth, from analysis. I, thereafter, calculated empirical BQ for each BQ score estimate (eq. 2.1):

$$\text{empirical BQ} = -10\log_{10}\left(\frac{\text{mismatch count}}{\text{match count}}\right) \tag{2.1}$$

To calculate the trinucleotide sequence context dependent CCS error rate, CCS reads from the cord blood sample were reconstructed, with the number of subreads for each CCS read set to 10 full-length subreads (the reasons are discussed in chapter 3). Cord blood CCS reads were subsequently processed as described above and below for read alignment and somatic mutation detection. To estimate the number of false positive mutations, the number of true positive somatic mutations were estimated from the number of callable bases and the cord blood somatic mutational process [101] and were subtracted from the number of trinucleotide sequence context normalised somatic mutation counts. The number of normalised false positive somatic mutation counts, and the number of callable trinucleotide bases were used to estimate the substitution and trinucleotide sequence context dependent CCS error rate.

2.2.4 Germline and somatic mutation detection

Germline and somatic mutations are both detected from bulk normal tissue leveraging CCS read length and base accuracy, characteristics unique to CCS reads and hard filters from previous publications [110, 111]. A BAM file with sorted primary read alignments is the only required input to obtain a VCF file with somatic mutations.

Upon initiation, read alignments are first randomly sampled from each target chromosome to compute the lower and upper bound read length and maximum read depth threshold $4d+\sqrt{d}$ where d is the average read depth. SBS candidates are collected from reads with average read accuracy, mapping quality score (MAPQ) and blast sequence identity greater than or equal to a predefined threshold. In addition, read length must be between the lower and upper bound read length to prevent somatic mutation detection from chimeric or fragmented reads. A naive Bayesian genotyper, thereafter, is applied to each SBS candidate to determine whether the data (D) only supports the variant as a germline mutation or whether the data support both a germline variant and a somatic mutation candidate simultaneously (eq. 2.2):

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)} \tag{2.2}$$

where $P(G)$ is the prior probability of observing the germline mutation genotype and D is the data that represents the pileup of read bases and corresponding sequencing error probabilities for each base at the substitution site. $P(D)$ is a constant across all the possible genotypes and is ignored. $P(G)$ is dependent on whether the genotype is heterozygous, heterozygous alternative (tri-allelic), homozygous alternative or homozygous reference allele with respect to the reference base (eq. 2.3):

$$P(G) = \begin{cases} \theta & \text{if } G = g_{\text{het}} \\ \frac{\theta}{2} & \text{if } G = g_{\text{hetalt}} \\ \theta^2 & \text{if } G = g_{\text{homalt}} \\ 1 - \frac{3\theta}{2} - \theta^2 & \text{if } G = g_{\text{homref}} \end{cases} \tag{2.3}$$

where θ is the expected germline SNP frequency and the default θ is set as 1×10^{-3}, the expected human germline SNP frequency.

$P(D|G)$ is the probability of observing the data given the genotype. Binomial likelihood is calculated for each genotype under the assumption that sequencing errors and read sampling is independent and identically distributed (eq. 2.4):

$$P(D|G) = \begin{cases} \frac{1}{2^n} \prod_i^n P(b_i|G) & \text{if } G = g_{\text{het}} \text{ or } g_{\text{hetalt}} \\ \prod_i^n P(b_i|G) & \text{if } G = g_{\text{homalt}} \text{ or } g_{\text{homref}} \end{cases} \tag{2.4}$$

where $P(b|G)$ is the probability of observing the base given the genotype and is defined as such (eq. 2.5)

$$P(b_i|G) = P(b_i|A) = \begin{cases} 1 - \epsilon_i & \text{if } b_i \in A \\ \frac{\epsilon_i}{3} & \text{if } b_i \notin A \end{cases} \tag{2.5}$$

where b is CCS base covering the target locus, ϵ is the corresponding sequencing error probability and A is allele of the genotype. In practice, all calculations are performed in log scale. Phred scaled likelihood (PL) is calculated for the 10 possible genotypes (AA, CA, CC, CT, GA, GC, GG, GT, TA, TT) using the posterior probability of the genotype (eq. 2.6):

$$\text{PL} = -10 \log_{10} P(G|D) \tag{2.6}$$

and PL for each genotype is normalised using the lowest PL (2.7).

$$\text{normalised PL} = [\text{PL}_i, \text{PL}_{i+1}, \ldots, \text{PL}_{10}] - \text{PL}_i \tag{2.7}$$

where PL is assumed to be sorted from the smallest to the largest. The genotype with the lowest PL is selected as the germline genotype. Genotype quality (GQ) score of the selected germline genotype is the difference between the second lowest normalised PL and the lowest normalised PL. If the data only provides evidence for a germline mutation, the next SBS is then considered for somatic mutation detection. If the data support the presence of both a germline mutation and a somatic mutation candidate, a number of conservative hard filters are subsequently applied to distinguish somatic mutations from errors:

1. If the germline mutation is a heterozygous, heterozygous alternative or homozygous alternative allele, somatic mutation candidate is excluded from the downstream analysis as somatic reversions are not considered. Somatic mutation detection, hence, is restricted to a locus with homozygous reference allele to prevent the misclassification of heterozygous mutation as a somatic mutation.

2. The GQ score for the homozygous reference allele needs to be above the minimum GQ score threshold.

3. The BQ score of the somatic mutation candidate needs to be above the minimum BQ score threshold.

4. Indels must be absent from the SBS locus.

5. The read depth of the target locus needs to be below the maximum depth threshold.

6. The reference allele count and the alternative allele count need to be above the minimum reference allele and alternative allele count. This condition is not required if the sample has sufficient sequence coverage as the GQ score is positively correlated with sequence coverage.

7. CCS reads with adapter sequences might still be present in the BAM file and misalignment of residual adapter sequences might generate somatic mutation candidates. Therefore, candidates located near start and ends of reads are filtered as specified with the minimum trimming parameter.

8. The number of mismatches adjacent to the candidate needs to be below the maximum mismatch count threshold within a given mismatch window as an alignment error can be mistaken as a somatic mutation.

A VCF file with common SNPs (>1% major allele frequencies) and a Panel of Normal (PoN) VCF file can also be optionally provided to exclude somatic mutation candidates potentially resulting from DNA contamination and systematic bioinformatics error, respectively. In addition, a VCF file with haplotype-phased hetSNPs can be provided to limit somatic mutation detection from haplotype phased CCS reads. Here, himut with default parameters (--min_qv 30 --min_sequence_identity 0.99 --min_gq 20 --min_bq 93 --min_trim 0.01 --min_ref_count 3 --min_alt_count 1 --min_hap_count 3 --mismatch_window 20 --max_mismatch_count 0) were used for the identification of unphased and haplotype phased somatic mutation. As sex chromosomes are enriched for misassembled regions and repetitive sequences [], somatic mutation detection was restricted to the autosomes. To process BAM, FASTA/Q and VCF files, himut internally uses pysam [112], pyfastx [113] and cyvcf2 [114], respectively. In addition, multiprocessing Python package [] was used to enable parallel processing of each chromosomes.

2.2.5 Panel of Normal construction

I created a PoN VCF file from 11 normal individuals with publicly available CCS datasets (Table X) to reduce the number of false positives arising from systematic bioinformatics errors. I ran himut with relaxed parameters (--min_mapq 30 --min_trim 0 --min_sequence_identity = 0.8 --min_hq_base_proportion 0.3 --min_alignment_proportion 0.5 --min_bq = 20) to maximise the number of mutations called from these samples. The number of samples in the PoN VCF is currently limited to the number of publicly available CCS dataset. As the number of CCS sequenced samples increases, the power to distinguish errors from somatic mutations will also increase in the future.

2.2.6 Germline mutation haplotype phasing

A haplotype is defined as a group of genetic variations that are inherited together from a single parent. I treat haplotype phasing as a graph algorithm problem where each hetSNP is a node in a graph and there is an edge between a pair of haplotype consistent hetSNPs. A single CCS read spans multiple heterozygous SNPs and evidence from multiple CCS reads can determine whether a pair of hetSNP is haplotype consistent ($p < 0.0001$, one-sided

binomial test) (Figure??). If a pair of hetSNP is haplotype consistent, a pair of hetSNP exists in cis configuration or trans configuration (Figure??). A haplotype inconsistent pair of hetSNP results from non-biological sources. Haplotype consistency is measured between all possible hetSNP pairs and hetSNP that is haplotype consistent with at least 20% of its possible pairs is connected through breadth-first search algorithm to construct contiguous haplotype blocks. Himut accepts as input VCF file with germline mutations and returns a VCF file with haplotype-phased hetSNPs.

2.2.7 Haplotype-phased somatic mutation detection

CCS reads are assigned to a haplotype block to enable haplotype-phased somatic mutation detection. To be allocated to a haplotype block, a CCS read must be within a haplotype block (and not between two haplotype blocks) and have haplotype identical to the consensus haplotype as defined in the haplotype block. In essence, somatic mutations are not phased through adjacent hetSNPs and instead phased CCS reads are used for somatic mutation detection. In addition, haplotype counts from the wild type CCS reads without the somatic mutation need to be above the minimum haplotype count threshold to select regions where both haplotypes have been sampled sufficiently and to prevent misclassification of hetSNPs as somatic mutations.

2.2.8 Somatic mutation count normalisation

To normalise the number of substitutions per trinucleotide sequence context, the SBS96 classification system and the same conditions as somatic mutation detection is used to the calculate of the number of callable reference and CCS bases. I would like to highlight that only the reference bases where homozygous reference allele has been called without an indel will considered as a callable reference base.

Under the SBS96 classification, SBS is categorised according to 6 possible substitution types in the pyrimidine context (C>A, C>G, C>T, T>A, T>C and T>G) and 16 possible trinucleotide sequence context derived from the 4 possible bases upstream and downstream of the substitution.

The frequency of each trinucleotide is calculated for the reference f_i^g, callable reference $f_i^{g_{\text{callable}}}$, and callable CCS f_i^{CCS} bases from the reference genome FASTA file, the number of callable reference bases and the number of callable CCS bases, respectively (eq. 2.8).

$$f_i = \frac{t_i}{\sum_{i=1}^{32} t_i} \tag{2.8}$$

where t denote a specific trinucleotide. There are 32 possible trinucleotide sequence contexts where pyrimidine is the middle base.

The number of somatic mutations is, thereafter, normalised with the ratio of reference to callable reference trinucleotide frequency and the ratio of callable reference and callable CCS base trinucleotide frequency according to the substitution and the trinucleotide sequence context. ACA>A somatic mutation count, for example, is normalised as follows (eq. 2.9):

$$S'_{\text{ACA>A}} = S_{\text{ACA>A}} \times r^{g}_{\text{ACA}} \times r^{\text{callable}}_{\text{ACA}} \tag{2.9}$$

where $S'_{\text{ACA>A}}$ is the normalised substitution count, $S_{\text{ACA>A}}$ is the raw substitution count, r^{g}_{ACA} is the ratio of reference to callable reference ACA frequency and $r^{\text{callable}}_{\text{ACA}}$ is the ratio of callable reference and CCS base ACA frequency.

The ratio of reference to callable reference trinucleotide frequency and the ratio of callable reference and CCS base trinucleotide frequency is calculated as follows (eq. 2.10):

$$\begin{aligned} r^{g}_{i} &= \frac{f_i^{g_{\text{callable}}}}{f_i^{g}} \\ r^{\text{callable}}_{i} &= \frac{f_i^{g_{\text{callable}}}}{f_i^{\text{CCS}_{\text{callable}}}} \end{aligned} \tag{2.10}$$

2.2.9 Mutation burden calculation

The somatic mutation rate is calculated for each trinucleotide sequence context from the normalised somatic mutation counts and the number of callable CCS bases (eq. 2.11)

$$m_{\text{ACA}} = \frac{S'_{\text{ACA>C}} + S'_{\text{ACA>G}} + S'_{\text{ACA>T}}}{t^{\text{CCS}_{\text{callable}}}_{\text{ACA}}} \tag{2.11}$$

where m_{ACA} is the somatic mutation rate for the ACA trinucleotide sequence context.

To calculate the mutation burden per cell, genomic mutation burden is first calculated using the trinucleotide sequence context specific somatic mutation rate and the number of trinucleotides in the reference FASTA file and the genomic mutation burden is adjusted with the ploidy of the sample to derive the mutation burden per cell (eq. 2.12)

$$g_{\text{burden}} = n * (\sum_{i=1}^{32} m_i * t_i^g) \tag{2.12}$$

where n is the ploidy of the sample, m_i is the trinucleotide sequence context somatic mutation rate and t_i^g is the number of trinucleotides in the reference genome.

2.2.10 CCS base quality recalibration

ACTC [] was used to align subreads to CCS reads from the same ZMW to determine the sequence orientation of subreads and to exclude subreads resulting from erroneous adapter sequence detection. DeepConsensus accepts as input the BAM file where subreads are aligned to CCS read from the same ZMW, polishes CCS reads and recalibrates the BQ score. CCS reads and subreads from the same ZMW were used to construct partial order alignment using abPOA [115] and the resulting alignment was processed to identify CCS bases where there is unanimous support from at least 10 subreads. CCS bases with unanimous support were assigned Q93 BQ score while all other bases were assigned Q0 BQ score. Himut was, thereafter, used to call somatic mutations from DeepConsensus polished CCS reads and CCS bases where there is unanimous support from at least 10 subreads.

2.3 Results

2.3.1 CCS library errors and sequencing errors

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

We generated 30-fold CCS sequence coverage from BC-1, HT-115 and blood granulocytes from an 82-year-old female individual (PD48473b) and 70-fold CCS sequence coverage from cord blood granulocyte (PD47269d) with an average read length between

16 and 20kb (Table 2.1) to achieve the following 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.

Table 2.1 Experimental Data

	BC-1	HT-115	PD47269d	PD4873b
Genomic DNA source	Cell line		Blood granulocyte	
Age (years)	-	-	0	82
CCS read count	5,962,252	5,933,281	12,156,251	4,949,180
Mean length ± std (bp)	18,571 ±	17,038 ±	16,523 ± 3,752	18,263 ± 1,753
Q93 bases (%)	51.4	55.5	57.6	51.7
Sequence coverage	36.9	33.7	67.0	30.1
Mutational process	APOBEC	POLE	Clock-like	
Mutational signature	SBS2	SBS10a, SBS10b and SBS28	SBS1 and SBS5	
Mutation burden per cell	~2,000 - 22,000	~8,000 - 11,000	~40 - 50	~1400 - 1500

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 processes 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 **??**) [].

CCS libraries are loaded on the SMRTcell and template DNA molecules diffuse 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 molecules 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 **??**). 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 **??**). 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 **??**).

CCS reads have an average read accuracy of at least Q20 and individual BQ score ranges from Q1 to nominal Q93, corresponding to 0.5×10^{-9} error rate (Figure **??**). 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 **??**) 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 **??**). 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 **??**). Using positive control samples, we identified additional CCS read characteristics that influences somatic mutation detection sensitivity and specificity.

2.3.2 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 **??**). Somatic mutation candidates are collected from CCS reads meeting the defined read-level prerequisites and candidates are categorised according to their base-level conditions (Figure **??**). 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 haplotype-phased somatic mutation detection (Method). Read-backed phasing with Illumina reads uses adjacent hetSNPs to phase approximately ~30% of detected somatic mutations []. In contrast, haplotype phased somatic mutation detection with CCS reads uses all hetSNPs that CCS read spans and phases approximately ~ 70% of somatic mutations (Figure **??**). 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**??**) 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 **??**), 3) CCS read with the same somatic mutations should share the haplotype and somatic mutations should not be present on both haplotypes (Figure **??**). Haplotype phased somatic mutation detection is especially helpful for samples with high heterozygosity.

2.3.3 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 the 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 [] as our somatic mutation detection method is agnostic to reference position.

We calculated mutation burden from BC-1, HT-115 and PD48473b samples to be X, X, X, respectively, consistent with previous estimates []. 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 samples (Method, Table). We also selected appropriate hard filter thresholds based on receiver operating

characteristic (ROC) curves generated under a range of hard filter conditions (Figure **??**) 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. The minimum BQ and GQ scores were crucial for somatic mutation detection while other filters had a 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.

2.3.4 CCS errors, error rate calculation and base quality score recalibration

The mutation burden in the cord blood sample is the lowest, with only 40-50 somatic mutations per cell []. CCS bases, unfortunately, do not have sufficient signal-to-noise ratio to enable somatic mutation detection in the 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 **??**). CCS error profile occurs in multiple samples, suggesting that the error process is systematic in nature (Figure **??**). 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 **??**).

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 [] and CCS reads with recalibrated BQ scores (Method, Figure **??**).

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 characters 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 **??**).

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 **??**), 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 identified 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 generates the expected mutational spectrum from the cord blood sample, suggesting that software error and not sequencing error is the source of false positive mutations. We hypothesise that the PacBio consensus sequence construction and polishing algorithm consider somatic mutations as errors and as a result have incorrect sequencing error priors and BQ score estimates.

2.4 Conclusion

Here, I assess whether CCS reads are as accurate as duplex reads and demonstrate that a subset of CCS bases has sufficient base accuracy to enable single molecule somatic mutation detection using samples with single ongoing somatic mutational process. Himut takes as input a sorted BAM file with primary read alignments from bulk normal tissue, leverages CCS read length and base accuracy to distinguish somatic mutations from errors and germline mutations and returns a VCF file with somatic mutations. Mutational spectrum produced from aggregate of somatic mutations is concordant with the expected mutational signature from each positive control sample, showing that single molecule somatic mutation detection is indeed possible with CCS reads.

Using a cord blood sample with few somatic mutations, I examined the nature of residual false positive substitutions and associated CCS error profile that is shared across all samples. I empirically estimated that CCS Q93 base accuracy ranges from Q60 to Q90 depending on the substitution and trinucleotide sequence context, which ~~is hundred thousand-fold to a billion-fold more accurate than~~ Illumina bases, and what enables somatic mutation detection with high confidence.

I conclude that false positive mutations are in fact derived from a combination of software errors. I show the persistence of inaccurate BQ score estimates using a modified pbccs that returns uncapped base quality scores, deepConsensus polished CCS reads and BQ score recalibration from partial order alignment between subreads and CCS reads from the same ZMW. I unexpectedly found that BQ score estimate becomes more inaccurate as the number of supporting subreads per CCS reads increases in contrast to the expected behaviour of the software (discussed and demonstrated in Chapter 3). In addition, I observe that false positive substitutions are enriched trinucleotide sequence contexts where the 5' base or the 3' base is identical to the substitution error. I hypothesize that inappropriate sequencing priors and underestimation of somatic mutations as potential sources of error in accurate BQ score estimation, and the use of trinucleotide sequence context HMM instead of dinucleotide sequence context HMM might ameliorate some of the issues. Most importantly, I show that subreads have sufficient base accuracy to generate CCS bases with ~Q90 base accuracy at all trinucleotide sequence contexts, if there is enough supporting subreads per CCS read.

2.5 Discussion

I conjecture that issue with CCS BQ score estimation will be properly addressed and that majority of CCS bases will have ~Q90 base accuracy in the imminent future. I, here, discuss the ramifications and potential applications following this development.

2.5.1 Somatic mutation detection

To date, CCS reads have been successfully used for construction of chromosome-length scaffolds of microbial and eukaryotic genomes [], used for germline SNP, indel and structural variation detection [], and have improved the genetic diagnosis rate of rare diseases []. The applications of CCS read for somatic mutation detection, however, have been limited and there has only been a handful of publications studying the complex structural rearrangements in cancers using CCS reads []. Here, I focused on single molecule somatic SBS detection with the intention to identify and analyse somatic mutational processes across the Tree of Life (discussed in Chapter 3) while others focused on improving the sensitivity and specificity of structural variations that could already be detected with Illumina reads []. Himut still cannot distinguish whether an individual SBS is an error or a

somatic mutation, but posterior probability can be calculated to determine the probability that the substitution is derived from a biological process or a non-biological process (??).

This approach previously was used to determine SBS16 mutational signature, a signature associated with alcohol consumption, as the main source of somatic mutations in CTNNB1 gene in hepatocellular carcinoma []. Despite this problem, himut will still enable researchers to rapidly screen for mutational signatures from bulk normal tissue without arduous experiments such as LCM or single-cell clone expansion sequencing, identification of environmental mutagenesis such as exposure to aristolochic acids[] across different locations and populations, lineage trace embryonic and tumour development through accurate detection of mosaic and somatic mutations, respectively. In addition, the ability to calculate the mutation burden in normal samples and thereby the age of the samples also raises the interesting question with regards to how to protect individual's privacy when SMRT platform becomes the primary sequencing method.

Himut currently does not consider matched tumour-normal sequencing for somatic mutation detection, but this would be the natural next step as the number of matched tumour-normal samples sequenced with the SMRT platform is expected to increase with the introduction of the Revio instrument. In the future, when error-free native DNA CCS library preparation is possible and when CCS BQ scores are correctly calibrated, HMW DNA extraction, input requirements for CCS library preparation and sequence coverage of the sample becomes the limiting factor to identifying and studying somatic mutagenesis across all tissues and all species.

In the interim, I believe that a wider range of somatic mutation detection will be possible with the benchmarking approach I have established where a sample with a known double base substitution and indel somatic mutational process is sequenced and used to fine-tune the pbccs algorithm and improve himut sensitivity and specificity. UV light, for example, induces the photoexcitation and dimerisation of adjacent pyrimidines into cyclobutane pyrimidine dimer (CPD) and 6-4 photoproduct. Although the exact mechanism that converts DNA damage to DNA mutation is unknown, CPD deamination has been suggested as one of the mechanisms generating C>T mutations (SBS7abc) and CC>TT mutations (DBS1) []. Cisplatin, a commonly used chemotherapy drug, forms interstrand DNA crosslinks to prevent DNA replication, which induces cell cycle arrest and apoptosis. Cisplatin produces a unique mutational signature where a single T insertion is introduced downstream of GG dinucleotides [], which is attributed to nucleotide excision repair of 1-3d(GpXpG) intra-strand cisplatin adducts []