**­­­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 \cite{Martinez-Jimenez2020-kn}, identify biological processes that generate these mutations \cite{Alexandrov2013-kg}, and to use somatic mutations as timestamps to lineage trace development \cite{Behjati2014-gb}. 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 \cite{Bailey2020-ou}. 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 \cite{Cibulskis2013-gw}. Library errors, introduced upstream of sequencing, is also often misclassified as somatic mutations \cite{Costello2013-cz, Chen2017-ba, Abascal2021-pk} .

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 \cite{}. 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 \cite{Li2008-dt}. 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 \cite{1000\_Genomes\_Project\_Consortium2012-rj}.

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 \cite{Lander2001-du}, 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 \cite{Osoegawa2001-np}. 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 \cite{Garrison2018-ae}. The Genome Reference Consortium (GRC) has released grch38 build to address some of these issues \cite{Schneider2017-yo} 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 \cite{Nurk2022-dv}. T2T assemblies, as expected, improve the accuracy and precision of both read alignment and variant calling \cite{Aganezov2022-dv}.

Table of current somatic mutation callers, their sensitivity and specificity, and their approaches

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 \cite{Lodato2018-hh}, single-cell clone expansion \cite{Lee-Six2018-qe} and laser-capture microdissection (LCM) \cite{Ellis2021-it} and sequencing adopts the former approach. 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 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 \cite{Schmitt2012-yr}. The duplex library is often diluted before PCR amplification to achieve optimal sampling and duplication per template molecule \cite{Hoang2016-jx, Abascal2021-pk}. 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) \cite{Schmitt2012-yr}

In contrast, duplex reads from the nanorate library protocol attains the promised Q90 base accuracy \cite{Abascal2021-pk}. 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 \cite{Chaisson2012-vr}. 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 \cite{Wenger2019-pw}. 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.

*Based on our understanding of duplex sequencing methods and the recently developed nanorate sequencing protocol, a derivative of the duplex sequencing protocol [ref, ref], and considering the similarities between two sequencing methods, we hypothesized that CCS reads might be as accurate or more accurate as duplex reads and can be used for single molecule somatic mutation calling. Both methods take advantage of the complementary information in the double-stranded DNA molecule to identify library errors and sequencing errors and to produce a highly accurate double-stranded consensus sequence (Fig 1). In the duplex sequencing protocol, a unique molecular identifier (UMI), consisting of 8 to 12 nucleotides, is added to the double-stranded DNA before the separation of the DNA molecule into single-strands for PCR amplification. Illumina adapters are added to the PCR amplified DNA, library is sequenced, and single strand reads belonging to the same DNA molecule is identified through their UMI. Because the forward and reverse strand is complementary and because DNA polymerase introduces errors infrequently during PCR amplification, library error and PCR amplification error should be present in one or two of the molecules, but not in all the molecules and most molecules should be consistent with each other such that the correct base can be ascertained. [bottleneck sequencing]. In contrast to the duplex sequencing protocol, CCS library preparation protocol doesn’t require a UMI as a single circular template is loaded to a zero-mode-waveguide (ZMW), DNA polymerase attaches to the template and sequences the circular template in a manner similar to rolling-circle-amplification to sequence both the forward and reverse strand multiple times until DNA polymerase terminates DNA synthesis [ref]. During DNA synthesis, DNA polymerase incorporates fluorescently labelled nucleotides and the fluorescence that is emitted during DNA synthesis is captured and fluorescence signal is converted to a nucleotide base. Thereafter, PacBio circular consensus sequence (pbccs) algorithm is used to generate CCS reads and for base quality score estimation.*

International efforts such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium \cite{ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts} and normal tissue sequencing studies from independent labs have sequenced thousands of genomes and have identified hundreds to thousands of somatic mutations per genome 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 \cite{Alexandrov2013-fq}; 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 \cite{Alexandrov2020-ys}. 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 \cite{Degasperi2022-qe}. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research \cite{Pich2019-ja, Aitken2020-sa}.

We invert the premise that long reads are inaccurate, demonstrate that CCS read is one of the most accurate sequencing platforms and discuss the ramifications following this observation.

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 \cite{Petljak2019-wi, Mitchell2022-ry}. 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.