**­­­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, mutation resulting from defective DNA damage repair, and the persistence of the mutation in the genome of the cell and its descendants (fixation of the mutation) \cite{}. 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 thousands of cancer genomes at scale and accompanied 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}, to use somatic mutations as timestamps to lineage trace development \cite{Behjati2014-gb}, to discover complex structural rearrangements such as chromothripsis that fundamentally changed the conventional view of tumorigenesis as the gradual process of the accumulation of somatic mutations \cite{} and to better understand the relationship between abnormal embryonic development and paediatric tumour formation \cite{}.

International efforts such as the Cancer Genome Atlas (TCGA) program \cite{} and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium \cite{ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts} have measured and analysed genetic, epigenetic, transcriptomic and proteomic aberrations in thousands of tumour genomes to understand how these aberrations contribute to the hallmarks of cancer \cite{} .

and have identified hundreds to thousands of somatic mutations

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 \textit{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). Double base substitution, indel and structural variation classification system also exist for mutational signature analysis, but they are not the subject of interest in this chapter \cite{}.

The PCAWG consortium has discovered 67 single-base-substitution (SBS) mutational signatures, and has determined the biological aetiology for 49 SBS mutational signatures (Table X) \cite{}. 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 sequenced genomes. Genomics England and collaborators, for example, have leveraged 100, 000 cancer 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}. Clinical sequencing of matched tumour and normal genomes is now routinely performed in the developed countries to help cancer patient treatment, fulfilling one of the many promises of the human genome project \cite{}.

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}. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure X). The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells \cite{}.

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{Lander2001-du}. If the repeat length is greater than the read length, read alignment software cannot determine the location of the read with respect to the reference genome 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 and variant calling is often restricted to the callable regions of the genome \cite{1000\_Genomes\_Project\_Consortium2012-rj}. Clinically relevant genes in non-callable regions, hence, are often excluded from analysis \cite{}.

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 and clone-by-clone sequencing and assembly of overlapping BAC clones is undoubtedly the best mammalian reference genome \cite{Lander2001-du}, but the human reference genome is still incomplete. The human reference genome, for example, still has missing sequences, unplaced scaffolds and unlocalised scaffolds without a reference coordinate, and misassemblies such as incorrect sequence collapse and expansion. Furthermore, 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 with alternative loci to address some of these issues \cite{Schneider2017-yo}. The recent completion of telomere-to-telomere CHM13 (T2T-CHM13) haploid genome using a combination of sequencing and mapping technologies has been a major milestone for genomics research \cite{Nurk2022-dv}. T2T-CHM13 genome, 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 \cite{}.

Illumina’s technical limitations have limited somatic mutation detection to clonal or sub-clonal mutations. 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 \cite{} and duplex sequencing (and its iterations) \cite{} adopt the latter approach where a highly accurate consensus sequence is created from multiple copies of a single molecule. 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. These methods have been successfully used for somatic mutation detection in normal tissues, including skin \cite{}, oesophagus \cite{} blood \cite{},

*including blood*[*6*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR6)*, placenta*[*7*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR7)*, neurons, smooth muscle*[*5*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR5)*, cardiac muscle*[*8*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR8)*, epithelia of the liver*[*9*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR9)*, bronchus*[*10*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR10)*, endometrium*[*11*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR11)*, colorectum*[*12*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR12)*, skin*[*13*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR13)*, esophagus*[*14*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR14)*, bladder*[*15*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR15)*, pancreas, prostate, ureter, thyroid, visceral fat, adrenal gland and testis*[*8*](https://www.nature.com/articles/s41588-022-01296-5#ref-CR8)*.*

These studies have informed on the clonal structure of tissues, somatic mutation rates, mutational processes and the presence of driver mutations conferring selection in normal cells of healthy individuals.

Duplex sequencing, however, is the most scalable for option for ultra-rare somatic mutation detection and is the preferred method for circulating tumour DNA (ctDNA) based clinical applications \cite{}.

Based on our understanding of duplex sequencing methods \cite{Schmitt2012-yr, Hoang2016-jx} and the recently developed nanorate sequencing protocol \cite{Abascal2021-pk}, a derivative of the duplex sequencing protocol and considering the similarities between two sequencing methods, we hypothesized 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 (Figure X). The duplex library preparation protocol starts with the sonication and fragmentation of genomic DNA. Unique molecular identifier (UMI) consisting of 8 to 12 nucleotide and Illumina adapters are attached 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}. PCR amplified library is sequenced and 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 generated leveraging the redundancies and complementarity between the forward and reverse strand reads. DNA polymerase, for example, might incorrectly replicate the template molecule during PCR amplification, but the polymerase 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 indicates the presence of polymerase 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 achieves 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. Moreover, the addition of dideoxynucleotides also inhibits nick translation, rendering DNA molecules that require this process unsuitable for library creation. The use of the restriction enzyme currently limits somatic mutation detection to where the recognition sites are present.

PacBio CCS sequencing also takes 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.

Here, we invert the premise that long reads are inaccurate and propose that CCS reads have the highest base accuracy among commercially available sequencing platforms.

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 and propose potential solutions to address these issues. In addition, we present himut, a method that can call somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome. 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 as the only input and returns a VCF file with somatic mutations as output. Our software is publicly available at <https://github.com/sjin09/himut> as a Python package under the MIT open license.

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 limit of detection, empirically calculate the PacBio CCS error rate and artefact signature resulting from systematic errors. 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}. The mutational spectra from previous studies and the contribution of different mutational signatures to the mutational spectrum serves 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.

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. Similarity between the mutational spectrum generated from the aggregate of somatic mutations and the expected mutational spectrum, hence, can be measured to assess the sensitivity of our method. In addition, mutational signature analysis can also be performed to determine the number of true negative and false negative somatic mutations from filtered somatic mutations to assess specificity.