**\section{Introduction}**

Based on my 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, 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. \newline

\textit{Most sequences have been derived by priming on both strands; this allows more confidence than when only one strand could be used} \cite{Sanger1977-os}.

\begin{flushright} [Frederick Sanger] \end{flushright}

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 \cite{Sanger1977-os}. To date, the Sanger sequencing method have been successfully used to obtain the 5,735 bp $\Phi$X174 genome sequence \cite{Sanger1977-os} and reference genomes sequences of \text{D. melanogaster}, \text{C. elegans}, and text\{H. sapiens} \cite{}.

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 the 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 \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 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 \ref{}). 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 strand is sampled sufficiently, complementarity between the two strands can be used to demarcate bases with high accuracy from bases with low accuracy \cite{Schmitt2012-yr} and to estimate the base accuracy from the supporting bases and associated base quality scores \cite{Abascal2021-pk}. Duplex reads, therefore, promises theoretical base accuracy of $1 \times 10^{-9}$ (Q90), but in practice, duplex reads from the original protocol achieves base accuracy of $1 \times 10^{-6}$ (Q60) \cite{Schmitt2012-yr}.

In contrast, duplex reads from the nanorate library protocol achieves the promised Q90 base accuracy and single-molecule resolution somatic mutation detection \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. 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 \ref{}). 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 \cite{Chaisson2012-vr}. CCS reads are reported to have an average read accuracy above Q20 \cite{Wenger2019-pw}, 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 \cite{} 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 ($\sim$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 \cite{}.

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 $\sim$17 somatic mutations per year per cell, equivalent to $\sim$1 somatic mutation per human genome per 6 weeks \cite{Mitchell2022-ry}. 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 \cite{Petljak2019-wi, Mitchell2022-ry}. 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 \cite{}. APOBEC mutational process inadvertently introduces C>T (SBS2) and C>G/C>A (SBS13) mutations to the genome at TCN trinucleotides (Figure \ref{}) \cite{} and localised hypermutations called kataegis, which are often observed at chromothriptic breakpoints \cite{}. APOBEC mutagenesis is, in fact, observed in more than 50\% of human cancers and accounts for considerable proportion of the total mutational burden \cite{}.

DNA polymerase $\alpha$ (POLA), $\delta$ (POLD) and $\varepsilon$ (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 \cite{}. 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 \cite{}. C>A mutations at TCN trinucleotides (SBS10a), C>A/C>T mutations at TCN trinucleotides (SBS10b) T>G mutations at NTT trinucleotides (SBS28) (Figure \ref{}) characterise POLE mutagenesis \cite{}.

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 \ref{} ) and cell division independent background mutational process (SBS5) (Figure \ref{}) \cite{} 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 \cite{}.