**\section{Materials and Methods}**

**\subsection{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 5% 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 (67563) 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.

***Sheared DNA was purified by solid-phase reversible immobilisation using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.***

***Pacific Biosciences HiFi circular consensus and 10X Genomics read cloud sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments.***

***All sequencing reactions were performed on the PacBio Sequel System with the Sequel Sequencing Kit 3.0 chemistry (Pacific Biosciences Ref. No. 101-500-400 and 101-427-800).***

***The HG002 human libraries were sequenced with 4 or 12 h preextension and 20, 24 or 30 h collection depending on insert length.***

**\subsection{CCS read alignment and germline mutation detection}**

CCS reads with adapter sequences were identified with HiFiAdapterFilt \cite{Sim2022-pi} 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) \cite{Li2018-am} and primary alignments were selected, compressed, merged, and sorted with samtools (version 1.6) \cite{Li2009-qp}. Germline SNPs and indels were detected with deepvariant (version 1.1.0) \cite{Poplin2018-ub}. VCF files were compressed and indexed with tabix \cite{Li2011-zj} and left aligned and normalised with bcftools (version 1.17-7-g097bda6) \cite{Li2011-ag}

**\subsection{CCS empirical base quality calculation}**

To assess the potential for somatic mutation detection with CCS reads, the accuracy of the BQ score estimate were assessed 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 \cite{Osorio2018-mh}, 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. We considered reference allele and germline SNPs as matches and all other SBS as mismatches. Germline mutation detection using himut is described below. We 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. We, thereafter, calculated Phred-scaled quality scores as such \ref{eq:1}:

\begin{equation} \label{eq:1}

\text{empirical BQ} = -10 \times \log\_{10} \Big( \frac{\text{mismatch count}}{\text{match count}} \Big)

\end{equation}

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 \cite{Mitchell2022-ry} and were subtracted from the number of trinucleotide sequence context normalised somatic mutation counts. The number of trinucleotide sequence context normalised false positive somatic mutation counts, and the number of callable trinucleotide bases were used to estimate the the trinucleotide sequence context dependent CCS error rate.

**\subsection{Germline and somatic mutation detection}**

Germline and somatic mutations are both detected from bulk normal tissue a without matched normal leveraging CCS read length and base accuracy, characteristics unique to CCS reads and applicable hard filters from previous publications \cite{DePristo2011-vf, Kim2018-qi}. 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 are collected from reads with average read accuracy, mapping quality score (MAPQ) and sequence identity above 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 to distinguish germline mutations from somatic mutation candidates.

To distinguish germline mutations from somatic mutations, a naive Bayesian genotyper is first applied to each substitution to determine whether the data $(D)$ supports a germline mutation or a somatic mutation candidate \ref{eq:2}:

\begin{equation} \label{eq:2}

\text{empirical BQ} = -10 \times \log\_{10} \Big( \frac{\text{mismatch count}}{\text{match count}} \Big)

\end{equation}

The genotyper considers 10 possible genotypes (AA, CA, CC, CT, GA, GC, GG, GT, TA, TT)

If a somatic mutation is detected,

heterozygous, heterozygous alternative (tri-allelic) homozygous alternative or homozygous reference allele.

Our method assumes that sample has a diploid genome. Our method first identifies the CCS read can be used for mutation detection

This step is done to discard reads that have large structural variations and that might originate from different genomic regions for mutation detection. Minimap2, for example, still has problems aligning reads with inversions. This step is done to restrict the mutation detection to reads where we are confident that the read has originated from the aligned region. Thereafter, single base substitutions, double base substitutions, multiple base substitutions, indels and complex variants are detected from each read.

To determine whether the detected single base substitution is a germline mutation or a somatic mutation detection, himut considers the 10 possible genotypes (AA, CA, CC, CT, GA, GC, GG, GT, TA, TT) and determines the most likely genotype based on the CCS bases and associated base quality score calculating the Bayesian binomial likelihood

[Eq XX, Eq XX] \cite{Li2011-ag}. In a normal tissue sample, the somatic mutation can occur on a homozygous reference, homozygous alternative, heterozygous or heterozygous alternative (tri-allelic sites) allele. We, however, do not consider the somatic reversion case where the homozygous alternative allele is reverted to the reference allele and ignore tri-allelic sites as the called somatic reversion can originate from genomic DNA contamination and tri-alleic sites account for 0.2\% of total known SNPs (ref, Heng LI).

P(D) is ignored as it is a constant across all the likelihood calculations.

We, hence, restrict the somatic SBS calls from bi-allelic homozygous reference sites as hetSNPs can also be misclassified as somatic mutation. We also require a minimum GQ score of 40 to have confidence that the site is homozygous reference, and the alternative allele must have a Q93 score for us to be confident that this is a somatic mutation and not a sequencing error. As incomplete adapter trimming is commonly observed in CCS reads, somatic mutations from the first 1\% and the last 1\% of the CCS read is ignored. In addition, if there is another mismatch within the defined mismatch window on the CCS read with the SBS, SBS is also discarded to avoid alignment errors being misclassified as a somatic mutation.

We assume that sequencing errors are independent and identically distributed to calculate the Bayesian binomial likelihood.

We have restricted the somatic mutation detection to autosomes as sex chromosomes often have lower quality assemblies and the repetitive content of the sex chromosomes causes more alignment errors.

In addition, VCF file with common SNPs (1\%>major allele frequencies) from public databases can be supplied to distinguish SBS arising from genomic DNA contamination. In addition, panel of normal VCF file constructed from himut with relaxed thresholds can be used to distinguish true SBS from that arising from systematic errors.

In addition, as reads originating from paralogous/orthologous sequences such as segmental duplications can align to off-target regions, SBS arising from sequence coverage above maximum depth threshold (4\*d + sqrt(d)) is discarded and SBS also needs to meet the minimum reference allele and alternative allele depth threshold.

Pysam \cite{pysam}, pyfastx \cite{Du2021-ya} and cyvcf2 \cite{Pedersen2017-ld} were used to process BAM, FASTA/Q and VCF files, respectively. In addition, multiprocessing Python package was used to enable parallel processing across multiple chromosomes.

**\subsection{Panel of Normal construction}**

We created a Panel of Normal (PoN) VCF file from 11 normal individuals with publicly available CCS dataset (Table X) to reduce number of false positives arising from systematic bioinformatics errors. We 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 maximize 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, in the future the power to distinguish somatic mutations from artefacts will also increase.

**\subsection{Germline mutation haplotype phasing}**

Haplotype phasing requires one to determine whether the polymorphisms are derived from a contiguous set of mutations. We treat haplotype phasing as a graph algorithms problem where each hetSNP is a node and measure haplotype consistency between a pair of hetSNPS to determine the validity of the edge. A single CCS read can span multiple heterozygous SNPs (hetSNPs) and a set of CCS reads can be used to measure the haplotype consistency between a pair of hetSNPs. Haplotype consistency if measured between all pairwise hetSNP and a pair of hetSNP is determined to be haplotype consistent through a binomial test (p<0.0001, one-sided). If a hetSNP is haplotype consistent with at least 20\% of its possible pairs, hetSNP is a haplotype consistent hetSNP. Using the breadth-first-search algorithm, haplotype consistent hetSNPS are connected to construct a haplotype block and both haplotype consistent and haplotype inconsistent hetSNPs are returned as a VCF file.

**\subsection{Haplotype phased somatic mutation detection}**

CCS reads are typically phased using adjacent hetSNPs. CCS reads, however, spans multiple hetSNPs and can be used to construct haplotype blocks. We use CCS reads to construct haplotype blocks (discussed below) and assign CCS reads to haplotype blocks. If the CCS read belongs to two haplotype blocks or if the hetSNPs belonging to the CCS read doesn’t match the haplotype phased hetSNPs exactly, CCS read is determined to be not phased. In addition, a hetSNP can be misclassified as a somatic mutation if the two haplotypes are sampled unevenly and hence we require both h0 and h1 haplotype counts of the wild type CCS reads without the somatic mutation in the region to be above the --min\\_hap\\_count 3.

**\subsection{CCS read base quality score estimation and recalibration}**

BAMsieve [ref, github] was used to select subreads where a productive ZMW created a CCS read with average read accuracy above Q20. abPOA \cite{Gao2021-nf} was used to construct partial order alignments between CCS and subreads from the same ZMW and the partial order alignments were parsed to select CCS bases where there was unanimous support from all the subread bases. The CCS bases with unanimous support was assigned Q93 base and all the other bases were assigned Q0 base and himut was used to call somatic mutations from CCS reads with recalibrated base quality scores.

XXX was used to align subreads to CCS reads from the same ZMW [ref, github] and samtools was used to compress the alignments and to select primary alignments. DeepConsensus (version --, command: ) \cite{Baid2022-or} takes as input the BAM file with subreads aligned to the CCS reads and returns polished CCS reads with recalibrated BQ scores. Himut was used to call somatic mutations from DeepConsensus polished CCS reads.

**\subsection{Single base substitution count normalisation}**

To determine the correct number of substitutions called per genome, the number of CCS bases where the substitution could have been detected from has to be determined considering the trinucleotide context frequencies in the reference genome.

\begin{equation}

f\_{i} = \frac{t\_{i}}{\sum^{32}\_{i=1} t\_{i}}

\end{equation}

\begin{equation}

r^{\text{callable}}\_{i} = \frac{f^{g\_{\text{callable}}}\_{i}}{f^{\text{CCS}\_{\text{callable}}}\_{i}}

\end{equation}

\begin{equation}

r^{g}\_{i} = \frac{f^{g\_{\text{callable}}}\_{i}}{f^{g}\_{i}}

\end{equation}

\begin{equation}

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

\end{equation}

\begin{equation}

m\_{\text{ACA}} = \frac{S'\_{\text{ACA>C}} + S'\_{\text{ACA>G}} + S'\_{\text{ACA>T}}}{t^{\text{CCS}\_{\text{callable}}}\_{\text{ACA}}}

\end{equation}

\begin{equation}

g\_{\text{burden}} = \sum^{32}\_{i=1} m\_{i} \* t^{g}\_{i}

\end{equation}

%\times r^{g}\_{\text{ATG}}

%\begin{equation}

%r^{g}\_{i} = \frac{f^{g\_{\text{callable}}}\_{i}}{f^{g}\_{i}}

%\end{equation}

We apply the same conditions as somatic mutation detection to all the CCS reads with and without the somatic mutation, determine the trinucleotide sequence context count from all the CCS bases where the same conditions would have been applied, calculate the ratio of trinucleotide sequence context frequency between the reference genome and the CCS bases. The single base substitution count is multiplied by the trinucleotide sequence context ratio to calculate the normalised single base substitution count. The normalised SBS count is used to calculate the mutation burden and to generate the mutational pattern plots.