**Chapter 2: Single molecule somatic single-base substitution detection**

**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, identify biological processes that generate these mutations, and to use somatic mutations as timestamps to lineage trace development [ref-ref]. 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 have different sensitivity and as a result, consensus somatic mutation call is often used for downstream analysis [ref, Nature Communications]. 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 [ref, MuTect2]. Library errors, introduced upstream of sequencing, is also often misclassified as somatic mutations [ref-ref, Science and NAR paper, Nanorate sequencing paper].

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. 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. 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 [ref, 1000G].

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 [ref, human genome project], 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 [ref]. 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 [ref, Ben, EKG, indel calling Rui]. The Genome Reference Consortium (GRC) has released grch38 build to address some of the issues. 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) for \textit{de novo} assembly and high-throughput chromatin conformation capture (Hi-C) reads for scaffolding [ref, ref, ref]. T2T assemblies, as expected, improve the accuracy and precision of both read alignment and variant calling [ref].

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, single-cell clone expansion and laser-capture microdissection (LCM) of clonal tissues and sequencing adopts the former approach [ref, ref, ref]. Rolling circle amplification, duplex and nanorate sequencing adopts the latter approach [ref, ref, ref].

Single-cell whole-genome amplification and sequencing, unfortunately, suffers from PCR amplification bias and allelic bias and hence, doesn't provide the most accurate somatic mutation calls and single-cell clone expansion and sequencing offers the most accurate somatic mutation call among the approaches that elevates the VAF of mutant DNA above the limit of detection threshold.

Duplex sequencing approach, inspired by the approach, that allows the use of DNA as a stable structure across millions of years uses the redundancy of complementary information between the forward and reverse strand of DNA double helix to achieve high base accuracy. In brief, DNA extract is fragmented through sonication, unique molecular identifier (UMI), consisting of 8 to 12 nucleotides (nt) is attached to a double-stranded DNA through blunt-end/overhang ligation with Illumina adapters, double-stranded DNA is separated into single-stranded DNA, which undergoes PCR amplification to produce multiple copies of the samme molecule. The concentration of input DNA for PCR amplification needs to be controlled to enable optimal selection and amplification of both strands of the double-stranded DNA. Illumina library is sequenced. Thereafter, reads are grouped according to their UMI and read belonging to the same UMI are used to generate double-strand consensus (duplex) read taking advantage of the complementary information between the forward and reverse strand. During PCR amplification, DNA polymerase might incorrectly replicate the template, but the incorrectly introduced base will be present only in one of the single-strand read or fraction of the single-strand reads and will not have a complementary base compared to the reverse complementary strand. Except for PCR-jackpot errors that occurs at the very first-cycle of PCR amplification and that impacts all the subsequent copies of the single-strand, duplex read should only be affecteted by library errors introduced upstream of PCR amplifiaction and sequencing. Duplex sequencing, theoretically, promises base accuracy of 1 x 10-9, but in practice duplex reads have a base accuracy of 1 x 10-6.

Nanorate sequencing improves the duplex library protocol to achieve the theoretical base accuracy of duplex reads. In contrast to the original duplex library protocol that repairs DNA molecules with DNA damage using end-repair enzymes, nick translation, and gap-filling enzymes, nanorate sequencing protocol prevents library preparation from DNA molecules with DNA damage. Instead of DNA fragmentation through sonication, nanorate sequencing protocol uses restriction enzymes that generates blunt-ends and through the addition of dideoxy nucleotides prevents DNA synthesis during nick translation. PacBio CCS sequencing shares many simliar features as duplex sequencing and we hypothesized that CCS reads might as accuraet or more accurate than duplex reads as CCS sequencing does not suffer from PCR amplifiaction and might have sufficient base accuracy to enable single molecule somatic mutation detection.

There are two types of third-generation sequencing: one from Pacific Biosciences and one from Oxford Nanopore Technologies and both companies attempt to sequence single molecule of DNA, in contrast to the sequencing by synthesis approach. These approaches, previsouly, had an error rate ranging from 20\% to 40\% depending on the library chemistry and the base caller version. PacBio introduced circular consensus sequencing in 2010, but circular consensus sequencing could not be adopted for mass-adoption as DNA polymerase for SMRT sequencing didn't have sufficient processivitiy to read long read-of-insert multiple times. Instead, PacBio offered continuous long read (CLR) sequencing to its customers which maximized for read length instead of average read accuracy. CLR reads typically have 10-15\% error rate, but is free from PCR amplification, the errors are thought to be randomly introduced and CLR reads have read length that is 100-fold longer than that from short reads. CLR reads, hence, was adopted for de novo assembly of complex genomes that could not be assembled with short reads and for structural variation detection. The longer read length enables the read alignment software to confidently assign the location of the reads relative to the reference genome as unique sequences are flanking repetitive sequences. Germline structural variation detection with long reads doubles the average number of structural variations discovered per genome compared to that from short reads and improves the diagnostic yield of rare genetic disease detection from 30\% to 80\%. The lower base accuracy and cost of SMRT sequencing, however, limited the wider adoption of PacBio SMRT sequencing except for one-off de novo assembly projects and clinical sequencing of patients with rare genetic diseases. PacBio, however, successfully engineered DNA polymerases with increased processivitiy and was further able to improve their circular consensus sequencing method such that read-of-insert with average read length of 10kb to 20kb can be read multiple times and because the errors are introduced randomly to each single-strand sequence templates, consensus sequence algorithms can take advantage of the complementary nature of double-stranded DNA to produce circular consensus sequences with average read accuracy greater than Q20.

Based on our understanding of duplex sequencing and the improved nanorate sequencing protocol, we hypothesized that PacBio CCS reads might have sufficient base accuracy for single molecule somatic mutation detection as it shares many of the characteristics as duplex sequencing, but is free from PCR-jackpot errors that occurs in the earliest stages of PCR amplification for duplex sequencing protocols. PacBio CCS base quality score ranges from Q1 to Q93, representing error rate from 1 to 1 in 5 billion bases. If the base quality score estimates are correct, single molecule somatic mutation detection should be possible in human samples as the human genome somatic mutation rate is 1 to 2 somatic mutations per human genome per 1-4 weeks, which is equivalent to Q70 and the somatic mutation rate is thought to be consistent throughout a person's life time. An individual has a higher somatic mutation rate than that of a normal person, if an individual has defective mutations in enzymes associated with DNA damage and repair process. If single molecule somatic mutation detection is successful from PacBio CCS reads, we aimed to detect and characterise the somatic mutations and somatic mutational processes in 66,000 eukaryotic species from the Darwin Tree of Life project, which aims to sequence and assemble all the eukaryotic species in Britain and Ireland, providing an unparalleld opportunity to examine the somatic mutations and assocaited somatic mutational processes of non-human samples across the Tree of Life (disscussed in Chapter 3). The study of somatic mutations across species allows us to tackel/attack the question posed by Peto's paradox. Why doesn't species with greater number of cells don't have higher incidence of cancer?

To date, we have sequenced hundreds and thousands of cancer genomes, we have been able to uncover the contribution of different somatic mutational processes responsible for the generation of somatic mutations in these samples through mutational signature analysis. In mutational signature analysis, mutations are classified into single-base-substitution (SBS), double-base substitutions (DBS) and indels. The somatic mutations are often presented in the pyrimidine context instead of purine context and multiple classification system exists for the single-base substitutions. SBS6 classification is used to classify the basic six types of substitutions (C>A, C>G, C>T, T>A, T>C and T>G). Using the 5' base and 3' base upstream of the substitution, 16 possible classification is for each substitution, creating 96 possible classification for the six substitution types. In addition, SBS96 classification system is sometimes decomposed to assess whether there is bias towards mutation occurence on the transcribed strand or the non-transcribed strand, creating SBS288 classification.

In addition, to our greater understanding of the role of somatic mutations in tumour evolution, we also have a better understanding of the biological processes that generates these somatic mutations. Mutational signature analysis [REF] across thousands of cancer genomes have

We need to define the problem first. In contrast to germline mutation detection where the mutation caller attempts to detect mutations that is homozygous or heterozygous, which exist as 100\% variant allele fraction and 50\% variant allele fraction, respectively, somatic mutations caller aims to detect somatic mutations that might be present in a single cell to a somatic mutations that might be present in all of the cancer cells and take into account tumour purity into the calculation. Somatic mutation callers, hence, often require a matched-tumour normal sequencing to distinguish germline mutations from somatic mutations and to calculate tumour purity??. In addition, because of techical limitations of short-read sequencing, low frequency somatic mutation with variant allele fraction below 0.1-1\% often cannot be detected.

To assess whether single molecule somatic mutation detection is possible with CCS reads, we sequenced three positive control samples (BC-1, HT-115, PD48473b) and one negative control sample (PD47269d, Table 1). The three positive control samples have a single somatic mutational process that is responsible for the generation of almost all the recently acquired somatic mutations in that sample. In contrast to the positive control sample, the cord blood sample should not have great number of somatic mutations and as a result, single-base substitutions detected from the negative control sample will be representative of the CCS error profile. In addition, the use of samples with single somatic mutational processes has the added benefit that these samples has been characterised in-depth through single-cell expansion and clone sequencing and we have determined the mutational probability of each substitution type in each trinucleotide sequence context. We, hence, are aware of the mutational pattern expected from the sample and can find the parameters that allows us to find mutational pattern from our positive control samples that is more consistent with what is expected from the sample. In addition, mutational signature analysis allows us to determine the number of mutations attributable to the correct biological process responsible for generating that somatic mutation and number of mutations attributable to false positive substitutions.

Our approach for single molecule somatic mutation detection is conceptually similar to using the genomic data from complete hydatidifiform mole 1 (CHM1) and CHM13 to assess the features important for heterozygous mutation detection. Because CHM1 and CHM13 cell lines have haploid genomes, the only valid germline mutations should be homozygous mutations and any detected heterozygous mutation must be an artifact of read alignment and germline mutation caller or sub-clonal somatic mutation that occured early in develompent. Similarly, as our samples have a single dominant somatic mutational process, any detected single molecule somatic mutation must be either a result of the biological process generating the somatic mutations or a result of library errors, alignment error or sequencing error, sytematic bioinformatics errors. Using the unique characteristics of our sample, we were able to find a search of features that impact the sensitivity and specificity of the somatic mutation calls. The sequencing statistics are summarised in Table 1. Here, we focused on single molecule somatic single-base substitution and the detection of larger structural variations that can only be detected with long-read sequencing is discussed in Chapter 4.

CCS read length and high base accuracy also allows us to haplotype phase the genome and call sub-clonal somatic mutations more effectively. If the detected mutation is a somatic mutation and not a sequencing error, the somatic mutation should be present on the same haplotype while that from sequencing error can be on both haplotypes.

To date, a number of approaches have been developed to detect somatic mutations using matched tumour-normal sequencing. Mutect..., Strelka2..., SomaticSniper..., Varscan2..., uses ... and Octopus..., to detect somatic mutations. To date, however, somatic mutation single-base substitution detection, have not been attempted with CCS reads and there has been limited effort in detecting somatic genomci rearrangements in cancer samples due to the availability of lack of scalable and reproducible methods.

Materials and Methods

Pacific Biosciences library preparation and circular consensus sequencing

%% indel errors

%% homopolymers

%% continuous long read sequencing

%% error profile have not been characterised

Library preparation

Standard protocol

Read alignment and germline mutation detection

CCS reads were aligned to the human reference genome (b37 and grch38) with minimap2 (version --) with the parameters “”[ref] and primary alignments were compressed, merged, and sorted with samtools (version --)[ref]. Single nucleotide polymorphisms and small insertions and deletions (indels) were detected with deepvariant 1.1.0 [ref].

Single molecule somatic mutation detection

himut accepts as input a sorted BAM file and returns a VCF file with somatic single-base substitutions.

{ATGC}

Germline mutation detection

Somatic somatic mutation detection

Heteterozygous mutation

HetAlt

Homozygous alt

Homozygous reference

\subsection{Mutation calling}

\subsection{Hard filters}

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.

CCS read base quality score recalibration

CCS read base quality scores

BAMsieve[ref]

Deepconsensus[ref]

CCS read and subread partial order alignment [ref] poatools

Results

We sequenced BC-1, HT-115 and PD48473b samples to approximately 30X sequence coverage and PD47269d to 90X sequence coverage.

%% BC-1, HT-115 cell lines

%% blood: SBS1, SBS5 + SBSX:

%% emily mitchell's paper

%% mia's paper

%% single dominant somatic mutational processes drive the generation of the most recently acquired somatic mutations

%% an individual/single substitution/difference between the read and reference genome cannot be determined to be a sequencing error or a true mutation that reflects the difference between the sample and the reference

%% if the circular consensus sequence base from Pacific Biosciences are sufficiently accurate, the differences between the read and the reference should be a reflection of biology and not the reflection of library errors, sequencing errors, systematic errors.

%% if single molecule somatic single-base substitutions, in aggregate, should be consistent with the expected mutational patterns from the sample if the mutations are correctly called.

%% sources of library errors, sonication, oxidative DNA damage

%%

%% single moelcule somatic single base substitution detection across all eukaryotic species, potentially sequenced with PacBio CCS sequencing method

%% even if soruces of errors introduced upstream of sequencing can be identified, it is hard to deal with them

%% systematic errors resulting from alignment errors and

%% Duplex sequencing

%% Nanorate-sequencing

%% Darwin Tree of Life project

%% Importance of somatic mutations: driver mutations, lineage trace development, time the emergence of driver mutations

%% somatic single base substitutions

%% small indels

%% structural variations > 50 bp

%% chromosomal translocations

%% tumour suppressor genes

%% technical limitations of illumina sequencing

%% normal sequencing: to understand the transformation of normal cells to neoplastic cells

%% DToL project

%% availability of reference genomes from diverse eukarytoic species

%% reference genome construction: CCS sequencing, linked reads, CLR, binano scaffolding

\section{Benchmarks}

\subsection{Sensitivity and Specificity, F1-statistics}

\subsection{Receiver-operating characteristics}

\section{Results}

\subsection{}

\subsection{}

\section{Discussion}

\subsection{Liquid Biopsy}

\subsection{False positive substitutions}

\subsection{Environmental mutagenesis}

has enabled us to detect more complex forms of genomic rearrangemetns such as chrmothripsis, chromoplexy, break-induced replication, fork stalling and template switching, etc and the elucidation of DNA damage and repair pathways that generate these complex rearrangements.