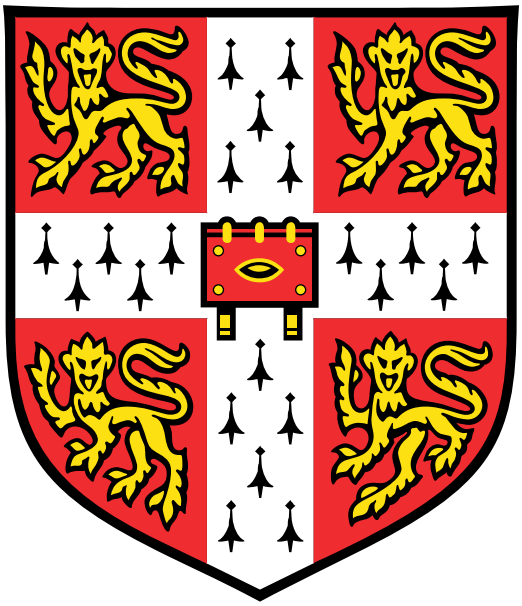
Single molecule mutation detection



Sangjin Lee

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College

January 2023

I would like to dedicate this thesis to my parents Kimok Lee and Misun Park
and my little sister, Hyunsong Lee.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Sangjin Lee
January 2023

Acknowledgements

"You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something - your gut, destiny, life, karma, whatever."

[Steve Jobs' 2005 Stanford Commencement Address]

This PhD thesis gives me the opportunity to reflect on my past and recognise the books, the events and people who have helped me to become who I am.

As a child, I was initially drawn to physicists with their acumen and ability to describe part of Nature with mathematics and later, I was inspired like many others to study the software of life and the manifestation of that software after reading What Is Life by Erwin Schrödinger. Three other books (Genentech: The Beginnings of Biotech by Sally Smith Hughes, Life at the Speed of Light: from the Double Helix to the Digital Life by J. Craig Venter and The Billion-Dollar Molecule: The Quest for the Perfect Drug by Barry Werth) also springs to my mind when I am asked which books inspired me to become a scientist. I don't know why, but I must have always loved the idea of a group of people working towards a shared goal to not only improve their understanding of the world, but to positively transform the lives of other people.

As an undergraduate studying biochemistry at Imperial College London, starting and finish a PhD degree was a distant dream and countless number of people have helped me achieve what I thought was impossible. My words cannot fully express my gratitude towards people who have helped me on my journey.

First, I would like to thank my parents. They have always believed in me. They have invested in my education. They have showered me with their care and attention. What I appreciate the most is that they did not ask me to conform to the social norms and instead they cultivated fierce independence to say no when it was necessary and to challenge and verify what I was taught and to have a healthy scepticism for everything. I could not have asked for a better family.

Second, I would like to thank Anny King, Rebecca Sawalmeh and Veronica McDouall for their care and warmth during my graduate studies at Churchill College. I still fondly remember weekly teak breaks with Anny, and light-hearted conversations with Rebecca. I absolutely could not have completed the MPhil in Computational Biology without their support. In the past, I dreaded waking up and I mightily struggled to complete the computational assignments. Now, I relish at the opportunity to design and implement new methods to explore the unexplored biological phenomena. How the tables have turned!

Third, I would like to thank Professor Jeong-Sun Seo, Chairman of Macrogen, for providing the opportunity to participate in the Korean Genome Project as part of my national service. I had no prior experience in sequence analysis, but he took a chance on me. I had the immense fortune to use the latest sequencing and genome mapping technologies to assemble chromosome-length scaffolds of the Korean reference genome. I cannot emphasize enough how important this research experience has been in increasing both my breadth and depth of knowledge and influencing the direction of research. Fourth, I would like to thank University of Cambridge and Wellcome Sanger Institute for the generous PhD studentship, creating an environment where I can be dedicated to research and providing the infrastructure to ask and answer original scientific questions. When I stroll through Cambridge, I am always in awe of the architecture and the fact I could breathe the same air and walk the same grounds as other great scientists who laid the foundation for human genomics.

Fifth, I have nothing but sincere gratitude towards my three supervisors Peter Campbell, Richard Durbin and Raheleh Rahbari for the opportunity to ask and answer original scientific questions. I had the unbelievable fortune to tackle three amazing questions: is genome-wide single molecule somatic single-base-substitution detection possible? If single molecule somatic mutation detection is possible, is single molecule structural rearrangement detection possible as well? What is the germline and somatic mutational process across the Tree of Life? I still cannot fathom the sequence of events that led me to this fortunate circumstance. I was the only PhD student in my year who was interested in exploring the capabilities and applications of PacBio circular consensus sequencing and Peter had the brilliant idea to assess the possibility of single molecule somatic mutation detection with PacBio CCS reads with samples with single ongoing somatic mutational process. An amazing opportunity presented itself and I was the only person who wanted to pursue it. I might not have another opportunity to work with such great supervisors and I wanted to record what I learnt and what I appreciated from them for perpetuity.

I think they believed more in me than I believed in myself and their confidence in me in turn motivated me to push myself and to burn the midnight oil. I cannot count the number of times I wondered if someone else might have been better suited to complete the projects. What I appreciated the most is that they had the courage to ask and attack the important questions and had the patience for me to make the mistakes and learn from mistakes such that I have ownership of my projects. I have been to many labs and I could not have had a better PhD and supervision elsewhere.

Sixth, I would like to thank my mentor for his wisdom and friends from high school (Anuran Makur, Gaurav Kankanhali, Jinseok Lee, Jisoo Kim, Kok Weng Chan and Victor Trisna), Imperial College (Claire Rebello, Euikon Jeong, Jiyae Kang, Jiyoon Kim, Jongseok Ahn, Quentin Godefroi, Rebecca Yu, Seonwook Park, Soo Young Yoon, William Gao, Woochan Hwang and Yunsung Na) and University of Cambridge (Dongseok Kim, Emily Sellman, Haerin Jang, Hans Werner, Hyesoo Lee, Ioana Olan, Ju An Park, Juyeon Heo, Kwon Juneyoung, Layla Hosseini-Gerami, Michal Tykac, Omid, Rob Henderson, So Yeon Kim, Sul Ki Park, Sunwoo Lee) for their continued friendship. Anuran and Gaurav have already completed their PhD and have started their assistant professorship at Purdue University and University of Pittsburgh, respectively. Jinseok just started his PhD at University of North Carolina at Chapel Hill and I have no doubt he will graduate with flying colours.

Seventh, I would also like to thank colleagues from Macrogen (Junsoo Kim, Chang-Uk Kim) and Wellcome Sanger Institute (Aleksandra Ivovic, Alex Cagan, Chiara Bortoluzzi, Chloe Pacyna, Emily Mitchell, Haynes Heaton, Hyunchul Jung, Jongeun Park, Jun Sung Park, Kenichi Yoshida, Lori Kregar, Matthew Young, Mike Spencer Chapman, Rashesh Sanghvi, Sigurgeir Olafsson, Thomas Mitchell, Thomas Oliver and Yichen Wang) for the stimulating conversations. A special mention goes to Mike Spencer Chapman and Heaton Haynes who were instrumental in maintaining my physical and mental health through regular afternoon runs and pair programming, respectively. If I have forgotten anyone in haste, you have my sincere apologies.

I will dearly miss my time at the University of Cambridge and Wellcome Sanger Institute.

Abstract

This is where you write your abstract ...

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Deoxyribonucleic acid (DNA)

"Let there be light", Genesis 1:3

Since the start of time, entropy has been increasing following the second law of thermodynamics and biological systems have emerged to reduce or maintain entropy using energy. Phospholipid permeable-membrane was the first spontaenous invention that separated order from disorder and allowed for the movement of molecules between the extracelular and intercellular environemtnt and for the emergence of primordial cell. It is uncertain whether the first cell had both the capacity to recplicate itself or whether had the capacity to catalyze chemical reaction first. In a prebiotic environment, amino acids can be created in a reducing environment if sufficient energy in the form of ioniznig radiation, ultra-violet light, is introduced into a gaseous atmosphere containing methane, ..., ... and ... [ref] and nucleotide bases are thought to be harder to spotaneously create in a prebiotic envrionment [ref]. Despite the uncertainity in how the first cell arised, the first prokaryotic organism is thought to have arise XX billion years ago and the first eukaryotic organism is thought to have arisen approximately 2 billion years ago [ref]. Once the first cell was created, selection pressure and natural selection acted upon these cells to create the first multicelullar organism and these muticellular organisms evolved to create multiple different species that is best adapated to the environment surrounding them. Mutations play a central role in creating new innovations that allows for individual species to better adapt to the environment and to produce progenitors that inherit the mutations.

It is now widely accepted truth that DNA is the unit of inheritance and that DNA has a double-helix structure and that the structure of the DNA drives many of the important

chemical reactions in the cells such DNA replication and transcription. In addition, sequencing technologies has become cheap enough such that clinical sequencing is routine enough to be able to detect the mutations that is responsible for disease and to understand the mutations that confer selective growth advantage to cancer genomes and amazingly, the cost of sequencing is still decreasing and new sequencing technologies are emerging to differentiate itself from short reads produced from next-generation sequencing platform. These widely accepted truth, however, were only enabled by giants who reimagined what was possible and who were willing to against the norm.

We must have wondered about the physical material that is responsible for the unit of inheritance from ancient times [ref]. [Greeks, Romans, Bible], Gregor Mendel is thought to be the father of modern genetics and provided the theoretical framework for the study of genetics with his famous experiment where the studied he inheritance of Peas's traits to their descenents in 1866X. Mendel carefully cross-breeded peas with different traits to discover that traits were inherited with a fixed ratio, also known as Mendelian ratio, and how certain traits are governed by dominant and recessive alelles. His experiment revealed how the physical material that is responsible for unit of inheritance must be separated into gametesand randomly united during fertilisation to determine the phenotype of the progenies and that the factors responsible for the phenotypic differences must be located independent of each other. These two rules are referred principle of segregation and principel of independent assortment.

Amino acids were initially proposed as the physical material responsible for inheritance as the number of amino acids and different varieties of protiens that could be created from different combinations of amino acould could potentialyl explain the complexity of a liviing organism and DNA was thought to be too simple to be able to encode the complexity of a living organism. It was not until the Oswald Avery's experiment in XXXX that demonstrated the DNA to be the physical material responsible for the transformation of R-strain bacteria to S-strain bacteria and despite the evidence, DNA was not believed to be physical material for unit of inheritance. The next race started with the aim of discoverying the sturcture of the DNA and there were many potential protagonists who could have discovered the structure of the DNA, but James Watson and Francis Crick, then post-doctoral fellow and PhD student at the laboratory of molecular biology, respectively, were the first to the race in 1954 [ref]. Despite the initial skepticism of how DNA could be the unit of inheritance and how DNA could be responsible for the complexity of an organism, the mechanisms of the central dogma was slowly revealed. Series of discoveries following the discovery of the structure of the DNA has cemented the importance of DNA

as the central unit responsible for directing cellular behaviours and determining phenotypes and encoding the software to produce proteins, the hardware that is responsible for catalyzing chemical reactions within the cell. Despite their simplicy, methods for DNA sequencing was designed later than that for amino acid sequencing. Frederick Sanger and Walter Gilbert cames with Sanger dideoxy sequencing and Maxam-Gilbert sequencing, respectively, to determine the nucleotide monomor that consistitutes the given nuclecic acid. Sanger was able to determine the genetic sequence of XXXX and XXXX using Sanger dideoxy sequencing for the first time. The Sanger dideoxy sequencing was more amenable to sequencing at scale and was adopted for the Human Genome Project (HGP) as the primary sequencing instrument and Sanger reads produced from ABI had an average read length of 500bp to 1000bp and had an average base accuracy between Q20 and Q50.

The Human Genome Project was initiated to sequence and assemble the human reference genome that would standardise the genetics and genomics studies to a single reference genome. There were two approaches towards the human reference genome construction: one was the hiearchical shotgun sequencing and assemgbly strategy and the other was whole-genome shotgun sequencing and assembly approach. The human reference genome constructed from the former approach is still the human reference genome used in most genetic and genomics studies and is the bedrock of genomic medicine revolution [ref]. The availability of the human reference genome together with sequencing-by-synthesis appraoch from Solexa, now Illumina, revolutionsed the field of human genetics and enabled population-scale studies of genetic diseases and cancers [ref]. Population-structure, human history, discovery of somatic mutations that confer selective growth advantage to the tumour cell, the identification of mutations that leads to genetic diseases. In addition, scientists have develooped clever ways to modify library protocol upstream of Illumina adapter ligation to enable the study of epigenomes, base modifiations, transcriptome of bulk tissue and more, recently, the advent of high-throughput chromatin conformation capture sequencing has enabled the study of the three-dimensional configuration of the genome and the nucleotide sequences are organised into regular repeating patterns.

The technical limitations of Illumina sequencing (base accuracy and short read length), however, has been the bottleneck for improving rare genetic disease diagnostics yield, detecting rare somatic mutations and constructing high-quality reference genomes for non-human species. De novo assembly of other species, previously, have been attempted using de brugjin graph based de novo assembly algorithms with short reads, but assemblies produced from short reads were highly fragmented and incomplete. In addition,

scaffolding strategies often did not provide sufficient long-range information to produce chromosome-level pseudomolecules and as a result, these assemblies provided incomplete information for comparative genomics purposes. Hence, assemblies produced from short reads often have collapsed repeats or contigs that cannot be placed accurately. To construct complete assemblies, reads need to be longer than the repeats of the target genome such that the reads can traverse the repetitive regions and optimally have unique sequences flanking the repetitive sequences such that the read can be placed in the assembly graph unambiguously. Not all repetitive sequences are actually repetitive. There are unique class of repeats called segmental duplications, which doesn't have a classical repetitive sequence, has a unique sequence, but is duplicated across the many parts of the genome and are thought to be important in driving evolution and these segmental duplications are typically defined as sections greater than 1kb with sequence simliarity above 90% to other regions of the genome. To distinguish segmental duplications from one another, reads also need to have high base accuracy to be able to distinguish closest segmental duplications from one another. Long-reads from third-generation sequencing technologies such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) provide an alternative towards improving the rare genetic diagnostics yield and improving the reference genome qualities in terms of both completeness and contigutiy. Long-reads produced from third-generation sequencing platforms were orders-of-magnitude longer than that from the Illumina platform, but had a much higher error rate; 10-15% error rate for continuous long reads (CLR) from PacBio and 20-35% error rate for ONT reads. Because of these high error rates, higher sequencing costs (lower yield per dollar) and insufficient improvement in read length, these platforms had limited use except for rare cases for real-time monitoring of ... and de novo assembly of plants and animal genomes..., and detection of pathogenic mutations that could not be detected with short reads [ref, ref]. Despite high error rate, the longer read length enabled the detection of structural variations that could not be previously detected with short reads, doubled the number of structural variations that can be detected from a typical human genome compared to the human reference genome. The longer read length allowed for the de novo assembly of BAC clones to hierarchically assemble missing sequences, also known as gaps, in the human reference genome, which have been problematic to assemble before and reveal human-specific gene duplications.

these companies have improved their library preparation protocol and base callers to improve the base accuracy. PacBio, for example, came up with circular consensus

sequencing protocl in 2014, but this protocol had limited use commercially until 2018 because of insufficient DNA polymerase processivitively.

1.2 Overview and objectives

Chapter 2

Single molecule somatic mutation detection

2.1 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 exhibit varying specificities and sensitivities. Consensus somatic mutation call, hence, 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 these 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) and high-throughput chromatin conformation capture (Hi-C) reads [ref, ref, ref]. T2T assemblies, as expected, improve the accuracy and precision of both read alignment and variant calling [ref].

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, single-cell clone expansion and laser-capture microdissection (LCM) and sequencing adopts the former approach [ref, ref, ref]. 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 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 [ref]. The duplex library is often diluted before PCR amplification to achieve optimal sampling and duplication per template molecule [ref, ref BotSeq, Nanorate sequencing]. 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) [ref, PNAS papers]

In contrast, duplex reads from the nanorate library protocol attains the promised Q90 base accuracy [ref]. 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. 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 [ref]. 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.

International efforts such as the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium and normal tissue sequencing studies from independent labs have sequenced thousands of genomes and have identified hundreds to thousands of somatic mutations per genome [ref-ref]. 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; 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 [ref]. 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 [ref, serena's paper]. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research [ref].

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 [ref, ref, Mia's, Henry's and Emily's paper]. 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.

2.2 Materials and Methods

2.2.1 CCS library preparation and sequencing

2.2.2 CCS 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]. Germline SNPs and indels were detected with deepvariant (version –).

2.2.3 Germline and somatic mutation detection

Our method first computes the average sequence coverage of the sample from random sampling of the read alignments across the genome to determine the average read length, read length standard deviation, sequence coverage and the maximum read depth threshold.

Our method assumes that sample has a diploid genome. Our method first identifies the CCS read can be used for mutation detection (-min_mapq 60 min sequence identity 0.99 –min_hq_base proportion 0.5 -min_alignment_proportion).

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]. 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, pyfastx and cyvcf2 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.

2.2.4 Panel of Normal construction

We obtained publicly available CCS reads (Table XX) and to maximize the number of systematic errors to be filtered, we relaxed the default thresholds (–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 construct a panel of normal VCF file.

2.2.5 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.

2.2.6 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.

2.2.7 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 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:) 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.

2.2.8 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. 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.

2.3 Results

2.3.1 CCS read characterisation

CCS reads have been successfully used for construction of highly contiguous and complete de novo assemblies and germline SNP, indel and structural variation detection for rare disease genetic diagnosis. In these applications, the accuracy of individual base quality

scores is not as important as 50% or 100% of the bases will support the consensus base, heterozygous or homozygous mutation. The accuracy of individual base quality scores, however, matters for ultra-rare somatic mutation detection as base accuracy must be higher than the human genome somatic mutation rate (1-2 mutations per 1-4 weeks per cell). Library, sequencing and systematic errors and genomic DNA contamination can be misclassified as somatic mutations, especially when a single read supports the alternative allele.

We generated 30-fold CCS sequence coverage from BC-1, HT-115 and blood granulocytes from an 82-year-old female individual (PD48473b) and 90-fold CCS sequence coverage from cord blood granulocyte (PD47269d) with an average read length of 15 to 20kb (Table 1, Figure XX) to achieve three objectives: 1), assess the potential for single molecule somatic mutation detection with PacBio CCS reads, 2) identify and address the sources of errors where possible and 3) empirically estimate the PacBio CCS error rate to define the limit of detection threshold.

To better understand the sources of sequencing errors, we first examined and identified sources of errors from CCS library preparation and sequencing. To create libraries with read-of-insert greater than 10kb, HMW DNA extraction is fundamental and is often carried out with either ..., ..., or Qiagen Magattract, or Circulomics HMW DNA extraction kit. If HMW DNA extraction is successful and if sufficient HMW DNA has been extracted, To create a topologically circulate template DNA, hairpin adapter is attached to the double-stranded DNA molecule (Figure X). DNA damage such as oxidative DNA damage introduced before or during library preparation is repaired using a cocktail of DNA repair enzymes (unpublished) and template DNA not suitable for sequencing is degraded using XXX DNase. The circular template, thereafter, is loaded to one of the ZMW in the SMRTcell and DNA polymerase at the bottom of the ZMW well initiates DNA synthesis using the circular template as a template. DNA polymerase incorporates fluorescently labelled free nucleotides, incorporation releases the fluorescent molecule, and the fluorescence is recorded through photonics and the wavelength of light emitted is recorded as one of the four nucleotide bases. DNA polymerase replicates the circular template through rolling circle amplification and sequencing terminates when DNA polymerase stops DNA synthesis. The DNA polymerase can initiate DNA synthesis from any starting points in the DNA template and equally terminate DNA synthesis from any point in the DNA template. Hence, the first and the last subread represents the partial readout of the template DNA while the second to the second subread are full pass subread that represents the full template DNA. DNA polymerase is agnostic to the strand orientation of

the template DNA and as a result, odd-numbered subreads and even-numbered subreads are assumed to have the same sequence orientation. The draft consensus sequence is constructed from multiple sequence alignment of subreads, and the draft consensus sequence is polished through the realignment of subreads to the draft consensus sequence. Dinucleotide sequence context Hidden Markov Model (personal communication with PacBio staff scientists) is used to infer the underlying DNA sequence (hidden state) and the base accuracy from the observed subread bases [ref]. The concordance of the supporting subread bases with the consensus base determines the CCS base quality score.

To better understand the CCS construction, subreads and CCS reads from the same CCS reads were analyzed together. We noticed that XX% of ZMWs have problems with adapter sequence detection, resulting in subread fragmentation and/or amalgamation (Figure XX); If the adapter sequence is incorrectly detected within the read-of-insert, the subreads can be split into multiple subreads and if the adapter sequence is not detected when present, two or more subreads can be connected to create a longer subread with both forward and reverse single-strand reads. CCS construction internally, hence, uses subreads that are longer than 50% of the median subread length and shorter than 200% of the median subread length. Despite this filter, full-length subreads are not purely selected and this filter doesn't account for ZMWs where adapter sequences are incorrectly detected in all the subreads. This phenomenon might explain CCS read that deviate from the read-of-insert length and these CCS reads that deviate from the read-of-insert length might be error prone.

We performed additional quality control to understand CCS performance (Figure XX). The cumulative proportion of the nucleotide bases should be consistent across the length of the reads, but the higher proportion of adenine and thymine at the 5' and 3' end of the CCS read is the result of A-tailing and incomplete adapter trimming.

PacBio also reports that as the number of subreads per CCS read increases, the average read accuracy also increases. We also confirmed that the increase in number of subread per CCS read also increases the number of differences as measured by the number of substitutions and indels per CCS read (Figure XX). Moreover, as the number of subreads increase per CCS read, the proportion of Q93 base also increases, but unexpectedly the bases are skewed towards Q93 bases and as PacBio supports BQ score ranging from 1 to 93, CCS reads also not easy to compress. The BQ score for CCS reads is capped at 93 as the ASCII standards cannot support higher scores and the user does not have access to the uncapped BQ scores. On average, DNA polymerase creates 10-16 subreads per CCS read per ZMW. The number of subreads per CCS read is a function of DNA polymerase

processivity, the rate at which DNA polymerase performs DNA replication and the read-of-insert length; The number of subreads per CCS read can either increase by increasing DNA polymerase processivity through protein engineering or by decreasing the read-of-insert length. The number of subreads and concordance between subread bases should be positively correlated with base accuracy. This, however, is not true in all circumstances and has unexpected negative ramifications as discussed in Chapter 3 and caution is required in choosing the read-of-insert length that will produce the CCS bases with the accurate BQ scores.

To date, CCS error profile has not been independently examined in depth

We initially used the positive control samples to assess whether Q93 CCS bases have sufficient base accuracy to enable single molecule somatic mutation detection and thereafter, used these samples to identify and assess features that influence sensitivity and specificity.

2.3.2 Germline mutation and somatic mutation detection with PacBio CCS reads

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.

The somatic mutation spectrum of a normal tissue is continuous as somatic mutation accumulation starts post-fertilisation and as cells with driver mutations expand and colonise greater proportion of the tissue and somatic mutation is an ongoing process resulting from intracellular and extracellular sources (Figure XX). Hence, genomic DNA extracts from normal tissue is a combination of DNA molecules that has germline mutations and somatic mutations. To distinguish somatic mutations from germline mutations in a tumour sample, matched tumour and normal sequencing is performed, but we are attempting to separate the germline mutations from somatic mutations in a normal tissue.

To distinguish germline mutations from somatic mutations, himut traverses read across the chromosomes to first find candidate single base substitutions from a set of CCS reads that meets a set of pre-determined alignment properties and thereafter, determines whether the single base substitution is a homozygous reference allele, homozygous alternative allele, heterozygous allele, or heterozygous alternative allele (tri-allelic sites) using a Bayesian classifier identical to that MAQ and GATK uses for germline mutation likelihood calculation (Methods). Once the germline mutation status of the reference position is

determined, himut only considers homozygous reference sites for SBS detection as other sites are candidates for somatic reversion and somatic reversions are not considered and somatic reversions might be the result of genomic DNA contamination. Himut, thereafter, applies a set of hard filters to mitigate the impact of the genomic DNA contamination and PacBio specific errors. To calculate the mutation burden of the sample, himut calculates the total number of trinucleotide sequence context that could have been potentially used for the somatic mutation calling with the same condition as somatic mutation calling and normalizes the mutation counts based on the trinucleotide sequence context frequency of the reference genome and callable bases (Methods). The user can prepare and supply a panel of normal VCF file to filter false positive somatic mutations resulting from systematic alignment errors and processing errors. In addition, true somatic mutations are haplotype consistent while false positive somatic mutations are haplotype inconsistent (Figure XX). To improve the sensitivity of sub-clonal somatic mutations, we take advantage of the CCS read length to haplotype phase the chromosome and use haplotype phased CCS reads for somatic mutation detection (Figure XX, Methods). Somatic mutation detection with short read sequencing uses adjacent hetSNPs to phase the somatic mutation and approximately 30% of somatic mutations are typically phased [ref, Serena's breast cancer paper]. In contrast, the longer read length allows haplotype phasing 70% of somatic mutations with CCS reads. In addition, to estimate the mutation burden of the sample, In the process of developing our method, we used the positive control samples to determine the features that are important for somatic mutation detection and suitable default parameters to be applied for future samples (Figure XX).

2.3.3 Somatic mutation detection sensitivity and specificity

Our method leverages the methods and approaches developed for germline and somatic mutation detection and improves upon them to apply our specific problem.

We applied our method to the positive control samples with different mutation burdens to obtain phased and unphased somatic mutations (Table 1). The mutation burden and mutational patterns from these samples were concordant to the mutation burden and signatures expected from these samples [Figure XX], demonstrating that PacBio CCS bases have sufficient base accuracy for single molecule somatic mutation detection. Using mutational signature analysis, we were able to determine the specificity and sensitivity of our method. Using mutational signature analysis, we can determine the number of true positive somatic mutations that fits the expected mutational signature of the sample and

what remains as the false positive somatic mutations; SBS2 signature is the only signature expected from the BC-1 sample and as a result, somatic mutations not attributable to SBS2 signature can be determined to be errors. Using the true negative, true positive, false negative and false positive somatic mutations, sensitivity, specificity and the F1 score of our method can be calculated. The number of true negative and false negative mutations can be determined from mutational signature analysis of filtered somatic mutations. We estimate himut to have XX%, XX% and X sensitivity, specificity, and F1-score, respectively. We, unfortunately, cannot compare himut with other existing somatic mutation callers as other callers are not designed for single molecule somatic mutation detection and/or somatic mutation detection is not technically possible.

The sensitivity improves from XX% and XX% and specificity increases from XX% to XX% when the grch38 human reference genome is used instead, reflecting that the higher quality assemblies leads to better variant calling.

In addition, we also assessed the impact of himut's individual parameters to sensitivity and sensitivity independent of other parameters while other parameters are maintained as a constant. As expected BQ and germline GQ score has the greatest impact on himut sensitivity and other parameters have small, but positive impact on sensitivity and the incremental additive effects of all the parameters in the resulting specificity and sensitivity (Figure XX). Moreover, we also assessed the sensitivity and specificity of each parameter thresholds and generated receiver-operating curve for each parameter to determine the best default parameter for somatic mutation detection (Figure XX).

In the process, we found artefactual mutational patterns that occurs consistently across all samples, which we refer to as CCS artefactual signatures. To determine the sources of errors that produces the artefactual mutational pattern, we examined the CCS and subreads together. As the artefactual signature appears in all samples, we hypothesized those upstream systematic errors must be responsible for generating these sequencing errors.

2.3.4 CCS error rate calculation and base quality score recalibration

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. The number of somatic mutations expected from the cord blood granulocytes are 40 – 50 somatic mutations per genome [reference Emily's paper and other papers]. Our colleagues have

also generated somatic mutations from single clone expansion and sequencing, the gold standard for single-cell somatic mutation detection and determined the ongoing mutational process in the cord blood granulocytes. The mutational pattern from cord blood granulocyte somatic mutations, unfortunately, was not concordant to what was expected from the sample, insinuating that the average CCS base accuracy is below Q93 as Q93 base should have been sufficient to capture all single molecule somatic mutations. We, however, used the false positive somatic mutations from cord blood granulocytes to determine the empirical CCS error rate. Using the cord blood HSC signature mutation probability and the trinucleotide sequence context count, we can estimate the number of somatic mutations expected from the sample, deduct that from the total called somatic mutations to calculate the number of mutations attributable to sequencing errors (Figure XX, Methods). We calculated the CCS base accuracy to range from Q60 to Q90 depending on the trinucleotide sequence context and the substitution (Figure XX, Methods)

We assumed that we have dealt sufficiently with the alignment errors and systematic errors in calling somatic mutation detection and wanted to determine the sources of errors upstream of germline and somatic mutation detection: library errors and sequencing errors. We did not focus on optimising the CCS library preparation to reduce the library errors as the Nanoseq protocol does to improve the duplex error rate. We, however, focused on identifying sources of sequencing errors. We hypothesized that CCS error rate must be resulting from incorrect CLR sequencing error priors. To test this hypothesis, partial order alignment between subread and CCS from the same ZMW was generated and we selected CCS bases with unanimous support from subread bases for somatic mutation calling (Methods). Somatic mutations called from CCS bases with unanimous support was concordant with what is expected across all the samples, suggesting that the inaccurate BQ score estimates are a software error and that this software error could be addressed with better subread substitution error priors. Google developed DeepConsensus to polish CCS reads with subreads and to re-calculate the BQ scores. DeepConsensus polished CCS reads have BQ score ranging from Q1 to Q50, and the estimates are too conservative compared our empirical estimations that can be derived (Figure XX). In addition, mutational pattern from Q50 somatic mutations is not concordant with what is expected from the sample, suggesting that the DeepConsensus polished CCS reads also don't have accurate BQ score estimates.

In addition, the use of samples with single somatic mutational processes has the added benefit that these samples have 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.

CCS BQ scores are capped at 93 as ASCII table doesn't support higher BQ scores. We collaborated with PacBio to obtain pbccs that returns uncapped BQ scores and observed the uncapped BQ scores for problematic trinucleotide sequence contexts where false positive substitutions are abundant are still a problem, suggesting that the base quality score needs to be recalibrated.

2.4 Conclusion

Here, we demonstrate that a subset of PacBio CCS has sufficient base accuracy to enable single molecule somatic SBS detection.

We estimate that CCS base accuracy ranges from Q60 to Q90 depending on the substitution and the trinucleotide sequence context. The CCS error rate is unexpectedly also dependent on the average number of supporting of subreads per CCS read (discussed in Chapter 3). The false positive substitutions resulting from inaccurate BQ scores are shared across samples and sequencing runs, suggesting that the issue is systematic in nature. Using a modified pbccs that returns uncapped BQ scores, we have confirmed that the same issue extends to CCS bases with BQ score above Q93. Google has developed deepConsensus to polish CCS bases and to revise CCS BQ scores based on multiple sequence alignments between subreads and CCS read from the same ZMW [ref]. deepConsensus BQ score estimates is capped at Q50, which is too conservative in comparison to our empirical calculation and similarly inaccurate as single molecule somatic mutation detection is not possible with deepConsensus Q50 CCS bases. We hypothesize the conservative deepConsensus BQ score estimate is since kmers arising from somatic mutations are treated as errors.

We observed that the false positive substitution is identical to the 5' and 3' and potentially the false positive substitution arises from the fact pbccs uses dinucleotide sequence context HMM and potentially a trinucleotide sequence context HMM might address the issue.

2.5 Discussion

To date CCS reads have ben successfully used for germline SNP, indel and structural variation detection and have improved the genetic diagnosis rate of previously undiagnosed rare diseases [ref, ref, Chaisson and Eichler, ngmlr, sniffles, deepvariant]. In addition, assemblies in combination with strand-seq enable detection of haplotype phased structural rearrangements longer than the read length [ref]. The applications of CCS read for somatic mutation detection, however, have been limited to date. Others have had limited success in using long reads for studying complex structural rearrangements in cancers and somatic retrotransposition detection [ref, ref]. The ability to detect large scale somatic structural rearrangements with long reads is especially important in determining the combination of genomic changes that results in the somatic structural variation. Here, we have focused on the successful detection of somatic SBS, but the method could be potentially improved to somatic indel detection. The somatic mutations detected from our approach are not all true somatic mutations and if a user wishes to determine the confidence of the somatic mutation call or determine the posterior probability of the somatic mutation call, user can calculate the posterior probability of the substitution coming from a specific trinucleotide sequence context to have been generated by a specific and known mutational signatures [,ref, Eq]. In the future, when the CCS base quality scores are properly calibrated, single molecule somatic mutation detection might be truly possible.

Here, we did not focus on identifying and addressing the CCS library errors. We, however, believe that library errors must be present in CCS reads. HMW DNA shearing using XXX, for example, introduces oxidative DNA damage. 5' filling or 3' filling with XXX enzymes can perform strand displacement and use the template strand to synthesize the complementary strand, and these processes have been documented to generate library errors (ref, Nanoseq). To eliminate the library errors, HMW DNA could potentially be obtained from blunt-end restriction enzyme digestion, perform A-tailing and hairpin adapters could be ligated through blunt-end ligase. In addition, DNA molecules dependent on strand displacement and synthesis can be made not-viable for library preparation with the addition of dideoxy nucleotides or with DNA restriction enzymes that digests single-strand DNA.

PacBio CCS bases are at least hundred thousand-fold to one million-fold more accurate than Illumina short read bases.

Our method and CCS sequencing can be used to identify the presence of MMR for immunotherapy purposes.

In addition, the method is focused on somatic mutation detection from normal tissues but can be extended to matched tumour and normal settings to enable sensitive somatic mutation detection from tumour tissues. We also attempted somatic DBS detection, which occurs in 100 fold less frequently than SBS, but like somatic SBS detection, true DBS signatures were outweighed by DBS artefact signatures.

We might be able to use a similar approach to also detect single molecule somatic structural variations.

During CCS sequencing, the kinetics of DNA polymerase during DNA synthesis is recorded. How fast, slow and whether the DNA polymerase paused during DNA synthesis is recorded. DNA polymerase kinetics data can be used to determine the base modification such as 5mC. Dennis Lo and colleagues, for example, have used ctDNA and NIPT DNA CCS reads to detect 5mC from single molecules and to successfully use them as diagnostic markers [ref]. Single molecule somatic mutation and 5mC together should provide greater sensitive with which tumours are classified, monitor their evolution and their potential trajectory under selection pressure.

HMW DNA input requirements for PacBio CCS reads limit the use of CCS sequencing for NIPT and ctDNA based genetic diagnosis (discussed in Chapter 5). HMW DNA input requirements are, however, expected to decrease with library preparation optimisation and like how DNA input requirements for Illumina sequencing has decreased.

Darwin Tree of Life project has sequenced and assembled high quality reference genomes using CCS and Hi-C reads, providing us with the opportunity to detect somatic mutations from other non-human samples, for the first time (discussed in Chapter 3). The somatic mutation rate and mutational signatures are unknown across these species. The study of somatic mutations across species allows us to tackle/attack the question posed by Peto's paradox: why doesn't species with greater number of cells don't have higher incidence of cancer?

We take advantage of the CCS base accuracy to detect gene conversions and crossovers in sperm samples and granulocytes from Bloom syndrome patients (discussed in Chapter 4). In addition,

PacBio has released new sequencing instrument Revio that increases the CCS read throughput 3 times with increase in read length and 3-fold increase in the number of ZMW, enabling the instrument to generate 30-fold sequence coverage genome at $1000. This should drive adoption and increase the number of human genomes sequenced with the PacBio instrument. Researchers will typically use CCS reads for de novo assembly or for germline structural variation detection, but collection of CCS reads from public

databases will enable the investigation of environmental mutagenesis across different populations across the globe and study the influence of germline mutation to somatic mutation generation and the combination of germline mutation and exogenous mutagen in generating new somatic mutagenesis.

The introduction of himut allows researchers to detect 5mC, germline SNP, indel and structural variation detection and somatic mutation detection from a single SMRTcell on the Revio instrument. The breadth and depth of sequence and epigenetic information provided by CCS reads compared to Illumina sequencing for a single run of sequencing at a single molecule level should enable better diagnosis and study of samples.

Three Matrix = Mutational signature probability

Mutational signature is itself an abstraction of the three steps of somatic mutation: DNA damage, incorrect DNA repair and fixation. The accuracy of the PacBIo CCS bases and the ability to detect 5mC might enable us to dissect/deabstract the SBS1 mutational signature. The spontaneous deamination of 5mC to thymine (C>T) at CpG site is detected and repaired by the MMR repair machinery. We know the mutation probability of the spontaneous deamination of 5mC biological process to generate somatic mutations at CpG contexts, but we are, however, unaware of the rate at which spontaneous deamination of 5mC happens in vivo and the rate at which the C>T substitution is repaired and unrepaired by the mismatch repair (MMR) machinery. Using the base accuracy and the ability to detect 5mC base modification, we should be able to determine the rates of in vivo 5mC, success probability of the MMR machinery and the rate at which the C>T substitutions are fixed in the genome. We can imagine a scenario where a specific region will have wild type reads with 5mC, but one of the reads will have a C>T substitution. The subreads that was used to construct the CCS read can be examined to see whether the deamination happened on one of the strands and whether the other strand has complementary GC bases with 5mC. We can use similar approaches in the future to examine the probability of mutagen to generate DNA damage, DNA repair fidelity and DNA fixation probabilities.

The application for our method abounds as our method can act as a replacement for many of the laborious processes that provide single-cell resolution somatic mutation calls. Our method cannot provide single-cell resolution somatic mutation calls, but we can provide through time-series sequencing of the same sample, the monitoring of the same somatic mutation to study the population dynamics of the sample. In addition, our method can be used to screen for ongoing mutational processes in the sample cheaply without needed to perform laborious single-cell clone expansion and sequencing.

Chapter 3

Germline and somatic mutational processes across the Tree of Life

3.1 Introduction

The Tree of Life encapsulates biological entities with 5 billion years of history on Earth, extinct species, survivors, and their descendants. The genomes of a select number of species, deemed biologically important, have been sequenced and assembled [ref,ref,ref, Drosophila melanogaster, C.elegans, Zebrafish, Mouse]. Homo sapiens, as a matter of fact, is one leaf in the Tree of Life and an unknown number of leaves remains to be studied. The completion of the human genome project and ramifications of the human genome project is undoubtedly a monumental moment in human genomics, but we far from studying and understanding the question "What is Life?" "What constitutes Life on Planet Earth".

Contracts, expands, fuses, inverts, rearranges, inserts, deletes, substitutes and copies and pastes, recombines, and the combination of all the above mechanisms to change the genome.

A number of factors has thwarted our efforts to understand species across the Tree of Life. These factors include the sequencing cost, read length, base accuracy, and computational costs, genome sequence complexity and ploidy.

The human genome project cost approximately 3 billion dollars, equivalent to dollar per base pair and required colossal effort requiring international collaboration across major sequencing institutions. Despite the gargantuan effort to physically map and assemble individual BAC clones, the human reference genome had missing sequences, unplaced and unlocalized scaffolds with unknown locations on the human reference genome. The p-arm of acrocentric chromosomes and centromeric sequences of every chromosome, for

example, remains unassembled because of their highly repetitive sequence content. The centromeric sequence in the latest human reference genome grch38, hence, is modelled and is not a true representative of the underlying sequence. In addition, the palindromic sequences in chromosome Y makes chromosome Y particularly difficult to assemble and the high degree of similarity between chromosome X and chromosome Y because of X-degenerate and X-transposed sequences. The human reference genome required the advent of new sequencing technologies with higher base accuracy and longer read length to correct misassemblies and minor errors and a new generation of human reference genome [ref, ref].

Segmental duplications defined as non-repetitive sequences with >90

The whole-genome shotgun sequencing approach and assembly approach with Illumina short reads might be scalable, but the assembly produced from this approach has been incomplete and uninformative and not suitable for population genetic analysis. The JCVI genome, for example, created from 500bp Sanger reads are devoid of segmental duplications.

The human reference genome is undoubtedly the most accurate mammalian reference genome and required a colossal effort to generate BAC clones, to determine the location of BAC clones through physical mapping and determining the BAC clones for minimal tiling path generation.

There were initially two competing approaches for human genome construction: whole-genome shotgun sequencing by JCVI and minimum tiling path by the NCBI?

The advent of PacBio CCS and ONT sequencing has been a game-changer/monumental/pivotal moment for de novo assembly and Hi-C based scaffolding has been a game changer for generating chromosome-length scaffolds. Hi-C reads, were originally used for interrogating the 3D structure of the genome, to understand how the genome is folded and tightly packed. Illumina mate-pair sequencing with different insert-sizes have been used for order and orienting contigs. Similarly, Hi-C reads can be thought of as mate-pair sequencing with read-insert sizes ranging from 100 to chromosome-length insert sizes. Hi-C reads were first used by XXX for scaffolding by XX. In the 3D space, sequences that are closer in linear space is also closer in 3D space and further linear distance, the further the sequences are also in 3D space. In other words, sequences that are in proximity are more likely to be together in 3D space and vice versa [ref]. In addition, DNA derived from the same chromosome are in more contact with each other. Chromosomes are isolated in 3D space. Using these features, assembled contigs can be clustered to chromosomes and contigs can be ordered and oriented. In addition, aberrant Hi-C read signals can be

used to detect misassemblies and to identify regions that needs to be separated. BioNano Genome mapping also has enabled high-throughput physical mapping of the genome at scale, but as sequence information is not provided by optical genome mapping and does not provide additional structural information that is different from chromosome-length scaffolds produced from contigs and Hi-C reads, long-read and Hi-C sequencing based de novo assembly is the method of choice for most large-scale de novo assembly projects. In addition, Hi-C contact matrix against the assembled chromosome-length scaffold can be manually inspected through visualisation to identify misassemblies and to correct misassembliess. In addition, the assembly graph constructed from pairwise read alignment can also be visualised to inspect problematic assembly regions.

These advances have enabled T2T-assembly of CHM13 haploid genome and T2T assembly of microbial genome can be routinely done with ease.

The ability to generate highly contiguous and highly complete chromosome-length scaffolds inspired many groups to revive genome assembly projects to revisit the problem of understanding our relatives in the Tree of Life. The Darwin Tree of Life project at the Wellcome Sanger Institute aims to sequence and assemble high-quality reference genomes for 66,000 eukaryotic species from *Britain and Island* with the most recent sequencing technologies. The DToL project initially used a combination of CLR, linked reads, BioNano genome maps and Hi-C reads to construct chromosome length scaffolds, but the combination of CCS and Hi-C reads have become the sequencing method of choice to construct reference genomes of different eukaryotic species. We hypothesized that our method for single molecule somatic mutation detection in human samples agnostic of clonality will be applicable towards somatic mutation detection across species agnostic of species. The understanding of somatic mutagenesis process in non-human samples have been limited to date and we thought this would be an opportunity to study both germline and somatic mutational in non-human samples, and in many species for the first time, to understand the evolutionary relationship of different mutational processes and the emergence and convergence of different mutational process across time.

This opportunity allows us to answer/address many questions that could not be addressed to date. This opportunity allows us to have an attack vector with which the question can be interrogated. What is the somatic mutation rate of different species? How has somatic mutation rate changed during the millions of years of evolution? Why do certain species don't have cancer? Why is there no relationship between the number of cells per species and the incidence of cancer for each species? How has other species evolved to

protect their genome integrity and how has DNA damage and repair mechanisms evolved to protect the DNA from hostile environment?

Relatives in the tree of life A subset of the branches in the trees of life Select number of leaves on the tree of life has been studied which in turn was limited by the sequencing cost and the technical limitations of the next-generation sequencing platform.

3.2 Materials and Methods

CCS library preparation and sequencing

De novo assembly, scaffolding and curation Darwin Tree of Life project members assembled, scaffolded, and curated the reference genomes. The specific method used is dependent on the species and the availability of data, but the method is similar across species. Contigs were generated using either hifiasm [ref] or hicanu [ref], and misassemblies were detected and purged with purgedup [ref]. If parental data was available, trio-canu was used to construct haplotype phased assemblies. The contigs were ordered and oriented using Hi-C reads and scaffolds were polished with Arrow to close gaps and to obtain a more accurate consensus sequence of the assembly. The chromosome-length scaffolds are, thereafter, manually inspected with Hi-C contact matrix to identify remaining misassemblies, to correct misassemblies and to scaffold contigs where there is sufficient Hi-C signal to connect, order and orient the remaining unplaced and unlocalised contigs. If RNA-seq or Isoform-seq was available, EBI gene annotation pipeline was used to obtain gene annotations from each reference genome. The de novo assembly is an ongoing process with improvements in sequence data and assembly algorithms and the method is subject to change with changes in availability of sequence data and assembly algorithms.

Phorcus lineatus preparation

To obtain the foot muscle of P. lineatus, the shell was cracked open and carefully the foot muscle was obtained. We dissected the foot muscle of the P.lineatus and sent the sample for HMW DNA extraction using circulomics HMW DNA extraction kit. Insufficient HMW DNA was obtained through shearing and hence, Blue Pippin size selection was performed to size select the library.

CCS read alignment and germline and somatic 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]. Germline SNPs and indels were de-

tected with deepvariant (version –) and germline hetSNPs were haplotype phased with himut (version 1.0.0). Somatic SBS were also identified with himut (version 1.0.0) with minor modifications to enable somatic mutation detection agnostic of species. Before somatic mutation detection, himut loads the deepvariant VCF file to calculate the germline heterozygosity prior and uses the prior for subsequent germline mutation detection and to distinguish germline mutation from somatic mutation.

HDP mutational signature extraction

Mutation signature analysis

3.3 Results

3.3.1 DToL project

The DToL project aims to sequence and assemble 2000 species in phase 1 of the project. To date, chromosome-length scaffolds of 600 eukaryotic species have been sequenced and assembled (Methods), of which number of species were CCS sequenced. The assemblies and the sequence data are publicly available. Thanks to the read length and base accuracy of CCS reads, contigs have a high contig N50 (Figure XX) and Hi-C reads enable the construction of chromosome-length scaffolds and the scaffold N50 is limited by the chromosome length. In addition, the assemblies typically have Q50-Q60 base accuracy, comparable to the base accuracy of the human reference genome [ref]. The assembly statistics for each species and the reference genome accession number is summarised in Table XX.

Of which XXX number of samples had diploid genomes. We excluded polyploid samples from the analysis.

3.3.2 Somatic mutation detection and evaluation

As the CCS read and the reference genome is derived from the same sample, homozygous mutations should be reflected in the reference genome, and any mutation detected must be either a heterozygous mutation, a somatic mutation, or an assembly error. In addition, As the CCS read and the reference genome is derived from the same sample, false positive substitutions originating from alignment errors should be significantly reduced.

Different samples have different heterozygosity and hetSNPs can be easily mistaken as somatic mutations. To confirm that our method is applicable to non-human samples, we obtained Phorcus lineatus samples with different ages (3 samples each from the 3-, 5-, 10-

and 15-year-old) to confirm the linear relationship between time and mutation burden per cell. We calculated the somatic mutation rate of Phorcus lineatus to be XXX per cell per year (Figure X). The tight bound on the linear relationship between time and mutation burden per cell gave us the confidence that our sample is applicable to all species.

The sequencing summary statistics for P. lineatus is summarised in Table XX. As the read-of-insert size decreases, the number of subreads per CCS read increases (Table). As the number of subreads per CCS read increases, CCS read should have higher proportion of CCS bases with Q93 bases. We, however, were aware from uncapped CCS BQ scores that increase in the number of supporting subreads does not necessarily lead to more accurate BQ scores. We sub-selected 10 full-length subreads from each productive ZMW and re-generated the CCS reads such that all the samples shared the same constant for comparison (Methods).

The age of the samples is unknown and hence, somatic mutation rate cannot be calculated per species basis, but we can make some reasonable assumptions based on the life cycle of the species in question to estimate the somatic mutation rate of each species. We have excluded insects that undergoes metamorphosis from the calculation of somatic mutation rate as the embryonic stem cells which grows into the larvae and adult cells are distinct and separated earlier in the life cycle of the insect [ref]. We identified XXX number of somatic mutations across XXX number of species and discovered X number of mutational signatures from the somatic mutations with unknown aetiology.

3.3.3 Mutational signature analysis

As the CCS read and the reference genome is derived from the same sample, homozygous mutations are assembly errors that were not polished, and heterozygous mutations are the true mutations (Methods).

We observed a high concordance between the germline and somatic mutational process, suggesting that the somatic mutational processes we discovered is an endogenous somatic mutational process much like the clock-like mutational process SBS1 and SBS5 in human samples. The detected somatic mutational signature could explain much of the germline mutational process in many of the species (Figure XX).

We found SBS1 and SBS5 mutational signature to be common in birds and mammals. We, however, also discovered SBSX in killer whales. Interestingly, SBS1 was not found in any other species while SBS5-like signature was commonly found in other species. We still do not know the aetiology of SBS5 and the presence of SBS5 in non-dividing somatic cells

suggesting that DNA replication is not the driver of SBS5 and SBS5 might be a composite of multiple different mutational processes [ref]. Our data suggest the combination of mutational process that produces SBS5 might be an ancestral one as it is shared by species separated by hundreds of millions of years of evolution.

In contrast, there were some species where the germline mutational process and somatic mutational process were distinct from one another. The pitfall of our experimental design is that only one sample is available from each species, and we can be confident of the identified mutational signature unless the mutational signature is observed in multiple species of the sample family or if multiple samples from the same species is sequenced. We hypothesized that environmental mutagenesis might be responsible for the observed mutational spectra as it has a strong transcriptional strand bias and a strong preference for a specific trinucleotide sequence context (Figure XX). To confirm that this environmental mutagenesis is common in this species, we collected and analysed a number of additional hoverflies (Figure XX) and compared the mutational spectrum of species where multiple samples are available (Figure XX). The species could be clustered based on the similarity of the mutational pattern observed in each species (Figure XX).

Germline mutational processes are typically studied in the context of TiTv ratio to measure the ratio of mutations that are purine mutations to pyrimidine mutations. Human germline mutations typically have a TiTv ratio of 2.0-2.1. If the mutation process was truly random, the TiTv ratio would be 0.5, but because spontaneous deamination of 5mC to thymine is the common germline mutational process in humans, transitions are more frequent than transversions. TiTv ratio, hence, can give us an indication of what might be the frequent germline mutational process in other species (Figure XX).

We studied the germline mutational process in the light of somatic mutational process that we discovered. To compare the two mutational processes, we compressed the SBS96 into SBS48 for comparison as the ancestral allele is unknown for the germline mutation while the ancestral allele is known for the somatic mutation. To comparison revealed that much of the germline mutational process can be explained by the detected somatic mutational process while the remaining germline mutational process might be originating from mutagenesis associated with recombination or other unknown factors in each individual.

In addition, we discovered new PacBio artefact signatures independent of that one discovered and discussed in Chapter 2. The discovered PacBio artefact signature, we believe to be from library errors.

3.3.4 Germline and somatic mutational processes

3.4 Conclusion

We discover XX number of mutational signatures previously undiscovered in previous studies and XX number of mutational signatures absent in database.

3.5 Discussion

We expected short-lived insects to have the highest somatic mutation rate, but in contrast to our assumption, many of the insects, especially insects belonging to the lepidtopera family has the lowest mutation burden. XX family which diverged from the lepidoptera family XXX mya ago, however, seems to experience increase in mutation burden with age. The difference between the two families is that while lepidoptera has a metamorphosis stage while XX family does not. In addition, the lepdioptera, coleptera, XX and XX that undergoes metamorphosis account for 80% of the insects, suggesting that insects that undergo metamorphosis has an evolutionary advantage against insects that does not. We conjecture that metamorphosis allows adult insects to have limited exposure to the DNA damage that might have accumulated during the larvae stage and that the imaginal disc that developed into the adult insect might be protected from DNA damage like the gametes in human samples [ref]. In addition, placenta is reported to have higher somatic mutation rate and higher number of chromosomal alterations as a tissue that is useful only for a limited amount of time [ref]. Similarly, caterpillars or young larvae stage of the insect might accrue more somatic mutations and chromosomal alterations. To confirm our hypothesis, gDNA from chrysalis and the adult insect of the same individual could be acquired and sequenced.

Chapter 4

Meiotic recombination

4.1 Introduction

4.1.1 Meiotic recombination

4.1.2 Haplotype Map

4.1.3 Methods to study meiotic recombinant products

Trio-sequencing

4.2 Material & Methods

4.3 Results

4.4 Discussion

Chapter 5

Conclusion and Discussion

5.1 Conclusion

In this PhD thesis, we challenge the preconception that PacBio CCS bases are inaccurate, and we claim that CCS bases are, in fact, sufficiently accurate for single molecule mutation detection.

To support this extraordinary claim, we accumulate extraordinary evidence to characterise the CCS sequencing process, identify sources of sequencing errors and empirically estimate the Q93 CCS base accuracy to between Q60 and Q90 depending on the substitution and the trinucleotide sequence context. CCS bases, hence, are a hundred thousand-fold to a million-fold more accurate than Illumina bases. In addition, we use samples with a single ongoing somatic mutational process to show that not only single molecule somatic mutation detection is possible, but also that the expected mutational pattern expected is directly observable from the called somatic mutations. Our approach is similar to how CHM1 and CHM13 cell-lines are used to assess heterozygous mutation calls can be used to assess and benchmark single molecule somatic mutation calls. DeepConsensus polished CCS reads, uncapped CCS BQ scores and CCS BQ score recalibration with partial order alignment between CCS and subreads from the same ZMW together indicate that pbccs assigns incorrect BQ score estimates, which is responsible for the false positive somatic mutation calls. We, here, have not explored whether library errors are a source of false positive substitutions, but we believe that CCS library preparation could be optimised to reduce library errors and further improve single molecule somatic mutation call sensitivity and specificity similar to how the Nanoseq protocol improves the duplex protocol to improve somatic mutation call sensitivity and specificity. Using our understanding, we develop and benchmark himut that enables single molecule somatic

mutation calls with PacBio CCS reads and himut is available as a Python package under MIT open license at https://github.com/sjin09/himut.git.

We have discussed the advantages and disadvantages of PacBio SMRT sequencing platform. Before the introduction of circular consensus sequencing, PacBio optimised for read length instead of base accuracy and offered continuous long read sequencing with average read length between 5kb and 20kb and error rate of 10-15%. CLR reads, hence, were limited to de novo assembly and germline structural variation detection. The advent of CCS reads, however, is a instrumental/monumental momenet in human genomics on multiple-levels. We never had a readout of genetic sequences at this accuracy at this scale with this level of base accuracy. CCS reads have an average read accuracy of Q20 and above, but CCS reads have base accuracy between 1 and 93 with a nominal error rate of 1 error per 5 billion bases. To date, there has not been an independent assessment of PacBio CCS base accuracy except for data described in this PhD thesis. We estimate the empirical error rate of Q93 CCS bases to be between Q60 and Q95 and the error rate is dependent on the substitution and the trinucleotide sequence context. In addition, PacBio has informed us that they use a dinucleotide sequence context hidden markov model for consensus sequence generation and base accuracy estimation, and the limited observation of sequence context might be responsible for the erroneous base accuracy estimation. Moreover, we were able to recover mutational pattern that was more consistent with the gold-standard mutational pattern from the sample when we recalibrated the base quality scores, providing further evidence that base quality scores are erroneously calculated for each base for each trinucleotide sequence context. It is unclear whether how the erroneous bases are introduced to the CCS reads and these erroneous bases must be introduced upstream of the sequencing process or be a result of systematic sequencing error, but a better consensus sequence algorithm will be able to address this problem in the future. We, furthermore, observed that somatic mutations called from shorter CCS reads have a higher number of false positive mutations than that called from longer CCS reads. Our hypothesis is that template with read-of-insert will have higher number of full passes and hence, more bases will be assigned Q93 base quality score, increasing the likelihood that erroenous library errors are assigned a high base quality score. In addition, we have observed in one of our sperm samples and in some of the DToL samples where Blue Pippin based size selection prior to CCS library preparation will introduce DNA damage to the template DNA such that C>T mutations are elevated in the overall mutation call. For a damage introduced upstream of CCS library preparation to have Q93, the DNA damage must be repaired such that the DNA base on both the forward

and reverse strand is erroneously repaired. We hypothesied that ** might be responsible for this type of erroneous DNA damage repair. Hence, a combination of library errors and consensus sequencing errors are present currently in the CCS reads. Since himut relies on base quality score as one of the features of single molecule somatic mutation calling, the increase in the proportion of bases with Q93 bases leads to distortions in the number of absolute number of called mutations and decreases sensitivity.

Previously, to detect gene conversions and crossover, a trio-sequencing was done or sperm-typing was done. Trio-sequencing, however, can only capture 1 meiotic event per chromosome per child while sperm-typing is restricted to a known hotspot. Our approach, however, assesses gene conversions and crossovers across the genome where there is sufficient sequence coverage and hetSNP density to haplotype phase the target region.

We tackled another original question to assess the genome-wide meiotic and mitotic recombination products in sperm samples and Bloom syndrome patient samples and compare and contrast characteristics of meiotic and mitotic recombination. Gene conversions and crossover detection requires long-range haplotype phasing of hetSNPs and individual reads to detect recombinant products that contains both maternal and paternal hetSNPs. The standard Illumina reads, unfortunately, cannot be used haplotype phase multiple hetSNPs at a time while CCS reads with their longer read length and is able to span multiple hetSNPs. CCS reads also have sufficient base accuracy to have confidence that the hetSNP flip is a result of not sequencing error, but a biological process. We successfully demonstrate that not only single molecule somatic single-base-substitution detection is possible, but also that single molecule gene conversion and crossover detection is possible with CCS reads. The detected gene conversion and crossovers are located on known meiotic recombination hotspots.

Our understanding of germline and somatic mutational processes of non-human species has been limited to date. The availability of both CCS reads and high-quality reference genomes from the Darwin Tree of Life project creates an opportunity to study both germline and somatic mutational processes. We used himut to call somatic mutations across the DToL eukaryotic species, discover XX number of mutational signatures, of which XX were distinct from known COSMIC mutational signatures, indicating the presence of distinct DNA damage and repair process operational in other species. In XX% of species, germline and somatic mutational process were analysed to be similar like how clock-like mutational processes (SBS1 and SBS5) are responsible for germline mutagenesis in sperms and oocytes. In addition, some of these endogenous somatic mutational processes were shared in insects, which are known to have diverged 450 million years ago (mya),

suggesting the mutational signature that we have discovered might be an ancient somatic mutational process or that these insects independently developed the same mutational process. Mother Nature, however, often doesn't change if there is an existing solution unless there is immense selection pressure and the author believes that the mutational process has been conserved across insects.

In XX% of species (hoverflies), however, germline mutational process and somatic mutational process were discordant and with strong transcription-bias, potentially suggesting environmental mutagenesis might be responsible for the observed somatic mutations. XX, XX, XX and XX insects undergo metamorphosis from caterpillar to adult insect and imaginal discs develop into adult insects. We, conjecture, that the absence of somatic mutations in some of the adult insects that undergo metamorphosis to the fact that larvae form and the adult insects are derived from independent embryonic stem cells. The adult insect is derived from the imaginal disc, which remains inactive under the metamorphosis in the chrysalis stage. Hence, somatic mutation that might have accumulated during the young larvae stage will not be passed on to the adult insect and the adult insect will be able to pass on their genome with limited DNA damage. The absence of somatic mutations in lepidoptera, however, might also be confounded with the short lifespan of the adult insects. It is interesting, however, that insects that undergo metamorphosis account for 80% of the insect population [ref] and there must have been a selective advantage to undergo metamorphosis despite the vulnerability that it might pose to the insect.

Wright's laws and Moore's law should enable PacBio to achieve economies of scale at an exponential speed and the future that we dream of might be closer than we anticipate.

5.2 Discussion

See things not as they are, but as they might be [J. Robert Oppenheimer]

Library errors, sequencing errors are absent and where input DNA requirement is not a constraint towards sequencing.

I imagine a future where we will be able to telomere-to-telomere sequence haplotype phased genome of a cell at a penny per cell and de novo assemblies are not required to infer the genome of the cell. In addition, the base accuracy will be so accurate that we can believe that every base is always representative of the underlying sequence.

Full-length Transcriptome and proteome per cell With base modifications

And where we will not be aligning reads to the reference genome for variant calling, but when we will be performing comparative genomics between the genome of a single cell

and that of the reference genome to study cellular heterogeneity and the collective impact on phenotype, wirings of a single cell, fine-tune the genotype to phenotype relationship and have a systematic engineering approach to understanding life across all species.

SMRT sequencing: the last DNA sequencing platform

"Nothing is more powerful than an idea whose time has come" [Victor Hugo]

Illumina platform was the sequencer of choice for most researchers and clinicians, and we were able to deliver the promise of genomics with continued decrease in compute, storage, and sequencing costs to greater and greater number of people. Illumina sequencing cost has decreased faster than Moore's law from XXX to XXX, but the rate at which sequencing cost has decreased had slowed in recent years (Figure XX). In addition, the read length and base accuracy of Illumina hasn't changed marginally, the only noticeable change/innovation has been in the throughput per lane. There is a limit to the knowledge that can be gained with marginal increase in number of genomes sequenced with Illumina sequencing platform. This is demonstrable from 30% rare genetic disease diagnosis rate with Illumina platform and the need to develop new protocols to study single-cell genomic and transcriptomic heterogeneity. And without competition, Illumina has not reduced their sequencing costs to maintain their profit and operating margin [Figure X]. We can conclude that for new technologies and new approaches are required to have a better understanding and to advance human genomics.

Third-generation sequencing or single molecule sequencing from ONT and PacBio was a hard sell for most consumers. The throughput was lower, error rate higher and sequencing costs was higher, and the read length was not substantially better than that from Illumina either. In the last decade, however, the both ONT and PacBio have substantially increased throughput, decreasing per base sequencing cost, and improved upon the base accuracy and the longer read length (>10kb-100kb) have started to interest scientists to revisit the problem of de novo assembly algorithms, structural variation detection and construction of high-quality plant and animal genomes. In addition, PacBio started to optimise their library preparation to optimise for read base accuracy instead of read length by increasing DNA polymerase processivity and keeping the read length constant.

The author, here, believes that PacBio SMRT platform could be the last DNA sequencing platform. The PacBio SMRT platform has the potential to be the cheapest and the most accurate and scalable sequencing platform in the market and PacBio has demonstrated excellence in execution and delivered on their promises. PacBio long reads have improved in base accuracy rate from Q10 to Q90 in the last decade, improved throughput CLR throughput from XXX to XXX and CCS throughput from XXX to XXX with the introduction

of Revio, which delivers whole-human genome at $1000, a competitive price considering that CCS reads can be used for de novo assembly, haplotype phasing, 5mC detection, somatic mutation detection and structural variation. (the versatile applications of CCS reads). Our research suggest that PacBio SMRT platform will be able to increase exponentially in the future as well with increase in the number of ZMWs per SMRTcell and increase in the read-of-insert-length. Our research also suggests that DNA polymerase processivity is no longer the bottleneck to obtaining Q90 bases and that CCS base quality score estimate is responsible for obtaining correct/incorrect BQ score estimates and hence, read-of-insert length can be further increased (Figure XX). The way in which the number of ZMWs per SMRTcell is increased is similar to how the number of transistors is increased per semiconductor chip and improvements in fabrications technologies from TSMC, ASML, Lam Research, Applied Materials have pushed the limits of what is possible. Furthermore, the acquisition of circulomics and optimization of CCS library preparation reduces the HMW DNA input requirements and in the future, we expect we can run SMRT sequencing from picograms of DNA. The trajectory of their improvement follows the improvements made on the Illumina platform (Figure XX).

The question is, hence, not whether PacBio SMRT platform is useful, but whether what will we do with reads produced from the PacBio SMRT platform.

The higher baser accuracy reduces the need to obtain higher sequence coverage to have the confidence with which the base is called.

In comparison to the traditional next-generation sequencing methods, CCS reads have longer read length, is free from PCR amplification and has higher base accuracy. Despite these limitations, PacBio CCS reads outperform on every metric from read length, base accuracy, number of applications from a single run compared to short reads from next-generation sequencing except for per base sequencing cost. This, however, is a limitation that PacBio as a company can overcome through a number of ways: i) the number of ZMWs per SMRTcell can be increased and ii) the average read-of-insert length can be incresed per template molecule. PacBio has increased the number of ZMWS per SMRTcell from XX ZMWs in XXXX to 8 million ZMWs to XXXX. In addition, the average read-of-insert length for CCS sequencing has increased from 10kb in 2019 to 20kb to 2021. Morever, if PacBio is further able to increase the processivitiy of DNA polymerase through further protein engineering or DNA polymerase evolution, they will be able to choose between longer average read-of-insert lnegth or increase in base accuracy through increases in the number of passes per template. I would assume that PacBio will choose to increase the read-of-insert length instead of base accuracy as base accuracy is

certaintly sufficiently high at the moment for most practical purposes and higher than what is offered through NGS platforms. In addition, our research suggests that PacBio CCS base accuracy problem should be resolved not through increase in the number of passes per read, but through better design of their conensus sequence algorithm. Recently, Google released deepConsensus algorithm to polish CCS reads based on alignment of subreads from the same ZMW to the CCS reads and to recalibrate the base quality scores. Deepconsensus, currently, cannot be applied towards all the CCS reads produced from SMRTcell and instead must be applied a subset of CCS reads for an average user. In addition, deepConsensus fails to estimate the base accuracy of the reads properly and the base accuracy estimates are too pessimistic, ranging from Q1 to Q50, which is below our empirical estimate between Q60 and Q90 for Q93 bases. In addition, if somatic mutations are called from CCS reads with polished with deepConsensus using Q50 bases, we are not able to obtain a mutational pattern that is expected from the sample.

Based on our understanding of CCS characteristics, we attempted to search for genomic events that could not be captured with short read sequencing and that could, however, be captured PacBio CCS sequencing. We hypothesised that PacBio CCS reads will also have sufficient base accuracy to detect gene conversions and crossovers from both sperm during meiotic recombination, granulocytes from Bloom syndrome patients and normal individuals during mitotic recombination. Gene conversion and crossover detection necessitates haplotype phasing of multiple kilobases and detection of haplotype rearrangement that might occur in a single sperm or a single cell.

Despite these limitations, as HMW DNA input requirements for CCS library preparation decreases and as sequence throughput and sequencing cost decreases, I believe that PacBio CCS sequencing might be the last DNA sequencing platform to dominate the sequencing market.

People don't have ideas. Ideas have people. [Carl Jung]

If we had the correct phylogenetic relationship between all species and mutational processes of all species on Earth, could we model and infer the mutational process of extinct species? Could we model and infer the mutational process of LUCA? Could we even derive the genome sequence of LUCA?

If life existed outside of Earth, what might be the mutational process responsible for speciation on other planets? How has Nature on other planets create new species? What is the creative process that Nature uses to create new species? Mutations are the paints that Nature uses to draw the Canvas.

We will be able to determine the ancestral mutational processes that shaped our genomes and the selection and evolution of mutational processes in light of different selection pressures that different environments applied our ancestors. As a consequence, we will also be able to determine the average fidelity of the DNA damage and repair process of all the species.

We don't know what might be the carrier of information that preserves the biological constraints of life might be on other planets.

The DToL project has sequenced 600 of 66,000 eukaryotic species in Britain and ... As the number

Kimura hypothesises that genetic drift would have been major driver of evolution and we would be happy to test this hypothesis.

The nucleotide composition of also extinct species. A thought experiment We are still early.

It might be possible to obtain sequence all of life within my lifetime and study/measure evolution in real time. Intelligence is equally distributed, and resources are unequally distributed. The unequal distribution of resources has been another factor that slows the understanding of all life on planet Earth.

During my bioinformatics career, PacBio has managed to improve their read base quality score a million-fold to a billion-fold while doubling the read length. In addition, what has traditionally required super-computers and international efforts to de novo assemble human genomes can now be done with a powerful laptop in a matter of hours thanks to new algorithms that makes the NP-hard problem de novo assembly problem to a more local problem that take advantage of the read length and base accuracy of the CCS reads and thanks to increase in the processing power of each semiconductor chip. The ability to cluster and phase reads based on their hetSNP and long-range information provided by Hi-C reads. We might be at the inflection point where we will be able to observe a Cambrian explosion in the number of new species studied.

We might be closer than we think on answering the question "What is Life" succinctly proposed by Erwin Schrodinger on XXXX at Dublin.

To have no stone unturned.

When the author whole-genome sequence analysis with Illumina reads, I cannot help but feel that I have not explored all that could be explored and that there might be something missing in the data that cannot be explored like the dark matter in the universe, which we know to exist, which we don't have any idea of its content. PacBio CCS reads resolves this issue.

References

Appendix A

How to install LaTeX

Windows OS

TeXLive package - full version

1. Download the TeXLive ISO (2.2GB) from
   https://www.tug.org/texlive/
2. Download WinCDEmu (if you don't have a virtual drive) from
   http://wincdemu.sysprogs.org/download/
3. To install Windows CD Emulator follow the instructions at
   http://wincdemu.sysprogs.org/tutorials/install/
4. Right click the iso and mount it using the WinCDEmu as shown in
   http://wincdemu.sysprogs.org/tutorials/mount/
5. Open your virtual drive and run setup.pl

or

Basic MikTeX - TeX distribution

1. Download Basic-MiKTeX(32bit or 64bit) from
   http://miktex.org/download
2. Run the installer

3. To add a new package go to Start » All Programs » MikTex » Maintenance (Admin) and choose Package Manager
4. Select or search for packages to install

TexStudio - TeX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads
2. Run the installer

Mac OS X

MacTeX - TeX distribution

1. Download the file from https://www.tug.org/mactex/
2. Extract and double click to run the installer. It does the entire configuration, sit back and relax.

TexStudio - TeX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads
2. Extract and Start

Unix/Linux

TeXLive - TeX distribution

Getting the distribution:

1. TexLive can be downloaded from http://www.tug.org/texlive/acquire-netinstall.html.
2. TexLive is provided by most operating system you can use (rpm,apt-get or yum) to get TexLive distributions

Installation

1. Mount the ISO file in the mnt directory

   ```
   mount -t iso9660 -o ro,loop,noauto /your/texlive####.iso /mnt
   ```

2. Install wget on your OS (use rpm, apt-get or yum install)

3. Run the installer script install-tl.

   ```
   cd /your/download/directory
   ./install-tl
   ```

4. Enter command 'i' for installation

5. Post-Installation configuration:
   http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-320003.4.1

6. Set the path for the directory of TexLive binaries in your .bashrc file

For 32bit OS

For Bourne-compatible shells such as bash, and using Intel x86 GNU/Linux and a default directory setup as an example, the file to edit might be

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/i386-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

For 64bit OS

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
```

```
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

Fedora/RedHat/CentOS:

```
sudo yum install texlive
sudo yum install psutils
```

SUSE:

```
sudo zypper install texlive
```

Debian/Ubuntu:

```
sudo apt-get install texlive texlive-latex-extra
sudo apt-get install psutils
```

Appendix B

Installing the CUED class file

LaTeX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TeXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LaTeXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LaTeXclass and package files. On some LaTeXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TeXLive systems this is accomplished via executing "texhash" as root. MIKTeXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LaTeX.