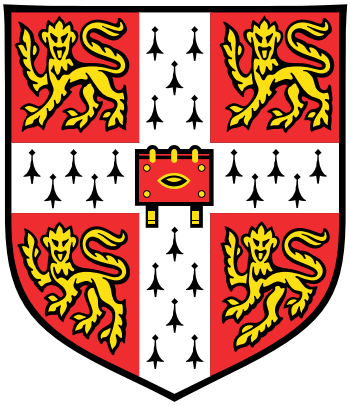
Germline and somatic mutational processes across the Tree of Life



Sangjin Lee

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College

March 2023

I dedicate this PhD thesis to my family who I have shared my hopes and dreams,
my joys and pains and my successes and failures.

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 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Sangjin Lee
March 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 Chuloh Yoon 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

Nomenclature

Acronyms / Abbreviations

5mC: 5-methylcytosine

BAC: Bacterial artificial chromosome

bp: Base pair

BQ: Base quality

CCS: Circular consensus sequence

CHM: Complete hydatidiform mole

chr: chromosome

CLR: Continuous long read

DNA: deoxyribonucleic acid

DNAP: DNA polymerase

dNTP: deoxynucleoside triphosphate

DToL: Darwin Tree of Life

gDNA: genomic DNA

GQ: Genotype quality

hetSNP: heterozygous single nucleotide polymorphism

Hi-C: High-throughput chromatin conformation capture

HMW: High molecular weight

indel: insertion and deletion

LINE: Long interspersed nuclear elements

mya: million years ago

OLC: Overlap layout consensus

ONT: Oxford Nanopore Technology

PacBio: Pacific Biosciences

PS: Phase set

SBS: Single-base-substitution

SINE: Short interspersed nuclear element

SMRT: Single-molecule real-time

SNP: Single nucleotide polymorphism

STR: Short tandem repeat

SV: Structural variation

TiTv: Transition to transversion

YAC: Yeast artificial chromosome

ZMW: Zero-mode waveguide

Chapter 1

from Last Universal Common Ancestor to the Darwin Tree of Life project

We cannot understate the significance that we can study physics and chemistry anywhere in the universe, but we can only study biology on planet Earth. We search for signs of life elsewhere in the universe, but we have yet to succeed in this endeavour. We must, hence, assume what distinguishes the inanimate from animate can be only understood here on Earth.

1.1 The Genomics Revolution

1.1.1 Genome assembly

Genome assembly aims to determine to entire genetic information of an organism. Genome assembly can be divided into four distinct stages: 1) shotgun or hierarchical shotgun sequencing and quality control to remove reads from contamination, 2) all-to-all read alignments to find overlaps between reads and to connect overlapping reads into contigs 3) to use long-range information to order and orient contigs into scaffolds, 4) and to assess and finish the genome through gap closing.

The ability to determine the nucleotide composition of organisms at scale with ABI capillary sequencing platform initiated a race to determine the genome sequence of scientific and economic interest and to determine the method that is most suitable for the human genome project. In principle, genome assembly aims to use randomly selected DNA fragments from the genome, to find overlaps between the DNA fragments and to connect the overlaps into a single contiguous sequence. If the genome in question does

not have repeats or if the read length is greater than repeat length, genome assembly becomes a trivial problem. Repeats account for less than X% of prokaryotic genomes. Repeats, however, are common in eukaryotic genomes and account for 50% of the human genome. Repeats take many forms and repeats can exist as tandem repeats, palindromes, or inverted repeats. There are repeats created by retrotransposons where retrotransposons use copy and paste mechanisms to create copies of themselves in the genome. Segmental duplications is a special type of repeat where non-repetitive sequences greater than 1kb with interchromosmal or intrachromosomal duplications with sequence identity greater than 99% []. Simple repeats such as short-tandem repeat (STR) expansions where dinucleotides or trinucleotides exist as tandem repeats.

In addition, These repeats create false overlaps between reads and these false overlaps either leads to misassemblies such as collapsed haplotypes or to disconnected contigs[]

Shotgun sequencing was initially used to create the first prokaryotic genomes of X, X and X and first eukaryotic genomes with Sanger sequencing. These genomes, thereafter, served as an excellent public resource to perform comparative genomics to find a common set of genes, to find conserved regions of the genome, to understand their evolutionary relationship.

1.1.2 Human Genome Project

Prior to the construction of the human reference genome through the Human Genome Project, the identification of pathogenic mutations in Mendelian diseases required the narrowing of the region with the likely causal gene through linkage analysis [], identifying the BAC clone that contains the sequence of the region through physical mapping, sequencing and assembling the BAC clone to retrieve the sequence of the region, and to find the pathogenic mutation through comparison with the BAC clone sequence [].

The availability of high-throughput Sanger sequencing instruments from ABI and initial success of construction of X, X, X and X genomes with Sanger reads inspired discussion to construct the human reference genome with aims to 1) accelerate the discovery of causal pathogenic mutations in Mendelian diseases 2), to create a single reference genome that can function as a single coordinate system for the scientific community to standardize research results, 3). Shotgun sequencing and hierarchical shotgun sequencing method were proposed for the construction of the human reference genome by JCVI and NIH, respectively []. Shotgun sequencing aims to assemble the genome from random DNA fragments sampled from the genome. Simulations has shown that if paired-end sequenc-

ing is performed on inserts of vary length with sufficient coverage, sufficient overlaps can be found to create contigs. In addition, mate-pairs can, thereafter, be used to order and orient contigs into scaffolds. Shotgun sequencing was proposed as an alternative to hierarchical shotgun sequencing approach as shotgun sequencing approach would not require the creation of BAC clones libraries, physical mapping of the BAC clones and independent sequencing and assembly of the BAC clones, thereby reducing the cost of the genome assembly drastically.

Prior to the completion of the human genome project, standardisation was absent from human genetic studies and the identification of pathogenic mutations in rare genetic diseases required arduous physical mapping and sequencing of BAC clones. The human genome project was initiated to determine the number of genes in the human genome, to accelerate the discovery of pathogenic mutations in rare genetic diseases, to expedite the drug discovery process. There were two competing efforts from the private sector and public sector with two distinct approaches to assemble the human genome. The private effort led by J. Craig Venter Institute (JCVI) used shotgun-sequencing approach and the public effort led by NIH used hierarchical shotgun-sequencing approach to assemble the human genome. Their contrasting aims led to differences in their methods. JCVI aimed to sequence and assemble the genome as fast as possible to patent the genes and to commercialize their proprietary database while the NIH aimed to create the most accurate human reference genome for biomedical research.

In contrast, NIH preferred hierarchical shotgun sequencing, also known as clone-by-clone, approach for construction of the human reference genome as the aims of NIH was not to create the assembly in shortest time, but to create a reference genome that can withstand the test of time and that can act as a focal point for scientific research and for scientific community. The hierarchical shotgun sequencing approach simplifies the assembly problem to the assembly of the 50-100kb BAC clone. Upon the successful assembly of the BAC clone, the location of the BAC contig can be determined from physical maps and overlapping BAC contigs can be assembled into a unitig []. Hierarchical shotgun sequencing approach aimed to use minimally overlapping BAC clones to create chromosome-length scaffolds for each contigs. The human genome project was an expensive enterprise and human reference genome is estimated to have cost 3 billion dollars. The human reference genome is undoubtedly one of the most accurate mammalian reference genome, but the human reference genome remains incomplete. The latest human reference genome build grch38 still has unplaced and unlocalized scaffolds and XX number of gaps, representing missing sequences []. The short arms of acrocentric

chromosomes are, for example, missing from the human reference genome. Unplaced and unlocalized are scaffolds where their location is not known and where their chromosomal origin is known, but their location is unknown, respectively. In addition, the centromeric sequences are not real and are modelled based on HuRef Sanger reads []. In addition, GigAssembler used for the Human Genome Project and Celera used for the HuRef assembly assumes that sequence data is derived from a haploid genome and if there is sufficient sequence divergence between two haplotypes in the same region, these assembly algorithms will collapse the two haplotypes into a chimeric haplotype that is not present in the population. Decoy sequences exist to prevent mismapping of sequences originating from satellite DNA to other regions of the genome and cause variant miscalling [].

The assembly quality was often assessed with paired-end reads from BAC clones. As the insert size and the expected orientation of the paired-end is known, if the insert size estimated from the paired-end read alignment and if the orientation of the reads are different from what is expected, these misoriented reads and misdistanced reads can be used to assess the assembly quality/scaffolding quality [].

Segmental duplications are often one of the common causes of genome misassemblies and where sequences are not successfully assembled resulting in missing sequences in the human reference genome[]. Segmental duplications have resulted in human-specific gene duplications not found in other great apes [], but these human-specific gene duplications are often missing from the human reference genome. Recovering these human specific gene duplications such as SRGAP2, NOTCH2L, BOLA2 required the selection, sequencing, and assembly of BAC clones to resolve these missing sequences. These human-specific genes have been associated with neocortex expansion and brain development [].

Updating and finishing the human reference genome is an ongoing process. The Genome Reference Consortium (GRC) is responsible for finding misassembled regions and updating the existing reference genomes of Homo sapiens, mouse, zebrafish, rat, and chicken. The update from grch37 to grch38 added X number of bases and was aimed to unify the existing different builds and was one of the first steps to better represent diverse haplotypes in different ethnic populations. The grch38 has X number of alternative loci and where each alternative loci represent a haplotype distinct from that in the human reference genome. GRC has used sequence data from CHM1 and CHM13 and CHM cell line BAC clones to resolve some of the existing issues in the human reference genome.

CHM cell lines are created when an egg without an embryo is fertilized with a sperm to create a cell line with a haploid genome [].

BAC clones were chosen as the vector of choice to retain large inserts as BAC clones were more stable than YAC clones and BAC clone DNA could be more easily amplified through E. coli culturing.

How is physical mapping done? Contamination removal

The human reference genome is continually updated to reflect the identification of misassemblies and to incorporate new sequencing and optical mapping data. The grch38 build, for example, currently has patch 13 with XX number of new bases [], but there is no immediate plans to release grch39 build. To better represent the genetic diversity and to improve variant calling sensitivity and specificity, genome graphs and variation graphs are under development to incorporate genetic polymorphisms into a graph and to provide a set of tools for scientific community to use the graphical representation of the reference genome for read alignment, variant calling, visualization [].

In addition, the advent of long and accurate single molecule sequencing technologies brings renaissance to the genomic assembly field (discussed later in the chapter).

1.1.3 Illumina sequencing

Somatic mutations can occur in cells at all stages of life and in all tissues. The biochemical manifestation of a somatic mutation requires three distinct stages: DNA damage or modification from either endogenous or exogenous sources, mutation resulting from incorrect DNA damage repair and unrepaired DNA damage, and the persistence of the mutation in the genome of the cell and its descendants [1]. Most somatic mutations are benign, but some confer a proliferative advantage and are referred to as driver mutations. The advent of next-generation sequencing and the continued decline in sequencing costs have enabled us to sequence thousands of cancer genomes at scale and subsequent downstream sequence analysis has allowed us to discover tissue-specific driver mutations [2], identify biological processes that generate these mutations [3], to use somatic mutations as timestamps and biological barcodes to lineage trace development [4], to discover complex structural rearrangements such as chromothripsis [5] that fundamentally changed the conventional view of tumorigenesis as the gradual process of the accumulation of somatic mutations [6, 7] and to better understand the relationship between abnormal embryonic development and paediatric tumour formation [8]. International efforts such as the Cancer Genome Atlas (TCGA) program [9] and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium [10] have also measured and analysed genetic, epige-

netic, transcriptomic and proteomic aberrations in thousands of tumour genomes to understand how these aberrations contribute to the hallmarks of cancer [11, 12].

Cancer is often described as the disease of the genome. Somatic mutation detection, hence, is often the first step towards characterising the cancer genome and these somatic mutations have been catalogued and analysed to determine their contribution to tumorigenesis. Multiple mutational processes simultaneously act on the genome at any given time and contribute 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 new mutational signatures or to assign the contribution of known mutational signatures to the mutation burden [13]; a mutational signature is a mathematical abstraction of the likelihood that a particular biological process will produce a somatic mutation in a specific sequence context. During mutational signature analysis, somatic mutations are classified according to the event, the size of the event and the sequence context. Single base substitutions (SBS), for example, can be classified using the SBS96 classification system, which categorises SBS according to the six types of substitutions in the pyrimidine context (C>A, C>G, C>T, T>A, T>C and T>G) and the 16 possible trinucleotide sequence contexts derived from the 4 possible bases upstream and downstream of the substitution. SBS can be further subclassified based on their pentanucleotide sequence context (SBS1536 classification) and whether the SBS is located on the intergenic DNA, transcribed or untranscribed strand of the gene (SBS288 classification). Double base substitution, indel and structural variation classification system also exist for mutational signature analysis, but they are not the subject of interest in this chapter [13–15].

The PCAWG consortium has discovered 67 single-base-substitution (SBS) mutational signatures [16]. To date, the biological aetiology for 49 SBS mutational signatures has been determined (Table X). The discovery of new somatic mutational signatures is an ongoing process where the number and the aetiology of mutational signatures is constantly updated and refined with increase in the number of sequenced genomes. Genomics England and collaborators, for example, have leveraged 100,000 cancer genomes from around 85,000 patients to detect mutational signatures associated with rare and sporadic somatic mutagenesis [17]. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research [18, 19]. Clinical sequencing of matched tumour and normal genomes is now routinely performed in the developed countries to help cancer patient treatment, fulfilling one of the many promises of the human genome project..

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 from multiple different detection algorithms, hence, is often used for downstream analysis [20]. The base accuracy and read length of Illumina reads, most importantly, is the common technical factor that limits 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 [21]. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations [22–24]. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure **??**). The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells.

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 [25]. If the repeat length is greater than the read length, read alignment software cannot determine the location of the read with respect to the reference genome as the read could have originated from any copies of the repetitive sequence [26]. The accurate placement of reads, hence, requires repetitive sequences to be flanked with unique sequences not present elsewhere in the reference genome. Consequently, the reference genome is divided into callable region and non-callable regions based on mappability of Illumina short reads and variant calling is often restricted to the callable regions of the genome [27]. Clinically relevant genes in non-callable regions, hence, are often excluded from analysis [28].

The completeness and contiguity of the reference genome is another often ignored, but important factor, for somatic mutation detection. The human reference genome constructed from physical mapping and clone-by-clone sequencing and assembly of overlapping BAC clones is undoubtedly the best mammalian reference genome [25], but the human reference genome is still incomplete. The human reference genome, for example, still has missing sequences, unplaced scaffolds and unlocalised scaffolds without a reference coordinate, and misassemblies such as incorrect sequence collapse and expansion. Furthermore, approximately 70% of the human reference genome is derived from genomic DNA of an anonymous individual of African-European ancestry [29].

The current linear sequence of the human reference genome, therefore, may not accurately reflect the genomic diversity present in other populations and alternative graph-based representations might better incorporate genomic diversity [30]. The Genome Reference Consortium (GRC) has released GRCh38 build with alternative loci to address some of these issues [31]. The recent completion of telomere-to-telomere CHM13 (T2T-CHM13) haploid genome using a combination of sequencing and mapping technologies has been a major milestone for genomics research [32]. T2T-CHM13 genome, as expected, improve the accuracy and precision of both read alignment and variant calling [33].

Table of current somatic mutation callers, their sensitivity and specificity, and their approaches [].

1.1.4 Challenges in somatic mutation detection

Cancer is often described as the disease of the genome. Somatic mutation detection, hence, is often the first step towards characterising the cancer genome and

these somatic mutations have been catalogued and analysed to determine their contribution to tumorigenesis.

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 from multiple different detection algorithms, hence, is often used for downstream analysis [20]. The base accuracy and read length of Illumina reads, most importantly, is the common technical factor that limits 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 [21]. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations [22–24]. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure **??**). The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells.

1.1.5 Single molecule somatic mutation detection

Illumina's technical limitations have limited somatic mutation detection to clonal or subclonal mutations. Two approaches have been developed to address these challenges: 1)

to increase the copy number of the mutant DNA above the limit of detection threshold and 2) to increase the base accuracy of the Illumina reads through upstream changes in the library preparation protocol. Single-cell whole-genome amplification [34], single-cell clone expansion [35] and laser-capture microdissection (LCM) [36] and sequencing adopts the former approach. Rolling circle amplification [37, 38] and duplex sequencing methods [39, 24, 40] adopt the latter approach where a highly accurate consensus sequence is created from multiple copies of a single molecule.

Single-cell clone expansion and LCM sequencing are recognized as the gold-standard methods for somatic mutation detection in single-cells or clonal tissues, respectively. These methods have enabled the study of embryogenesis, somatic mutation rate, mutational processes, clonal structure, driver mutation landscape and earliest transformation of normal cells to neoplastic cells across a range of normal tissues, including adrenal gland, blood, bladder, bronchus, cardiac muscle, colon, endometrium, oesophagus, pancreas, placenta, prostate, skin, smooth muscle, testis, thyroid, ureter, visceral fat [35, 41–56]. Duplex sequencing, however, is the most scalable option for ultra-rare somatic mutation detection and is the preferred method for circulating tumour DNA (ctDNA) based clinical applications [57].

1.2 Single molecule sequencing

Single molecule sequencing technologies from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) are spearheading the next decade of the genomics revolution. These upcoming technologies promise a new era of genomics where: 1) lower input material is required for library preparation and sequencing, 2) library preparation is location-agnostic and does not require skilled technicians, 3) sequencing takes hours and not days, 4) higher base accuracy, 5) longer read length (10kb – 100kb), 6) simultaneous detection of genetic variations and base modifications and 7) nucleotide-resolution identification of structural variations where event size is ≥50bp. Despite these promising capabilities, the higher error rate and marginally longer read length of first generation of ONT reads and PacBio continuous long reads (CLR) limited their adoption. Illumina is still the primary sequencing method in most labs as per base sequencing cost is still cheaper with the Illumina platform. After decades of development, ONT and PacBio have introduced new sequencing instruments and library preparation that exceeds the capabilities of Illumina platform in read length and accuracy, enabling researchers to

access the inaccessible regions of the genome and to explore biological phenomena that could not be explored before.

1.2.1 Nanopore sequencing

Cells use membrane proteins to move ions and molecules, critical to the maintenance of cellular function, across the permeable plasma membrane through passive and active transport []. David Deamer and George Church independently hypothesised 197X that a single strand of DNA molecule could be passed through a protein pore if voltage is applied through the membrane holding protein pore cite. If electrostatic potential is present across the protein. The disruption of the passage of ionic currents by the passage of the DNA molecule through the pore can be recorded and can be interpreted as a specific nucleotide base. Nanopore based sequencing methods promised 1) minimal library preparation, 2) ultra-fast native DNA and RNA sequencing and 3) unlimited read length

Today, Oxford Nanopore Technologies (ONT) has fulfilled many of these promises. To fulfil these promises, Deamer and colleagues had to demonstrate the potential of the Nanopore sequencing method through successive demonstrations and improvements of the technology that first shows that passage of the DNA through a pore and disruption of the ionic current is a detectable event [], and that a single nucleotide difference can be detected from a background of homopolymer sequence []. In addition, the first generation of pores based on $alpha$ had a pore that was too long such that 10-15 nucleotides will be interpreted as a single signal and hence, a pore that had a similar aperture, but shorter pore was required to improve the signal-to-noise ratio. MytA protein, hence, thereafter, was used for nanopore sequencing to improve the signal-to-noise ratio. To improve their base accuracy, ONT introduced 2D reads where both forward and reverser strand of the double-stranded DNA with a hairpin adapter could be sequenced through the nanopore []. ONT, however, long no supports 2D reads as a result of legal dispute with PacBio [].

ONT licensed these patents to commercialise the technology in 2005 and the most recent ONT reads are reported to have Q20 read accuracy []. To date, ONT reads have been successfully used to identify and characterise complex pathogenic mutations[], accelerate clinical diagnosis [], and to help the assembly of the complex regions in the human reference genome []. It could be said that ONT sequencing has fulfilled all of it promises and more.

1.2.2 Pacific Biosciences Single-Molecule Real-Time sequencing

PacBio was founded in 2004 with aspirations to commercialise single molecule real time (SMRT) sequencing technology developed at Cornell University. The SMRT platform is the culmination of multiple technical innovations from a range of disciplines. The zero-mode-waveguide (ZMW), a nano-scaled hole fabricated in a metal film, for example, is at the heart of the SMRT platform. The ZMW acts as the sequencing unit and its unique properties help the SMRT platform achieve the high signal-to-noise ratio required to observe activity of individual DNA polymerases (DNAP)[58].

The metal film with the ZMW is placed on top of a glass and DNAP is immobilised at the bottom glass surface through surface chemistry modifications that prevent the adsorption of DNAP to the metal side walls[59, 60]. A topologically circular template, also known as a SMRTbell template, is created through the attachment of hairpin adapters to a double-stranded DNA molecule (Figure X). The successful loading of SMRTbell template into a ZMW follows a Poisson distribution and typically 30 to 50% of the ZMWs are classified as productive ZMWs where a single DNAP successfully initiates and completes rolling circle amplification. SMRT sequencing initially used Φ29 DNAP for its high processivity, minimal amplification bias and ability to perform strand displacement DNA synthesis [60]. In addition, Φ29 DNAP was engineered through site-directed mutagenesis to use fluorophore-labelled deoxyribonucleoside triphosphate (dNTP) during DNA elongation [61, 60].

Upon successful loading of SMRTbell templates, free nucleotides are released above the ZMW array and free nucleotides diffuse in and out of the ZMW. DNAP binds and incorporates the correct nucleotide into the growing DNA strand, and upon nucleotide incorporation, DNAP cleaves the fluorophore from the nucleotide such that the synthesised DNA molecule consists of native DNA molecules. DNAP continues DNA elongation until DNA replication is terminated. The length of the reaction time is dependent on DNAP processivity and the presence of bulky DNA damage on the template DNA that can lead to premature termination of replication[]. Illumination from the laser below the glass surface excites the fluorophore and the emitted fluorescence is measured. Image processor leverages the temporal difference between diffusion of free nucleotides (which occurs in microseconds) and nucleotide incorporation (which occurs in milliseconds) to separate the background fluorescence from free nucleotides and fluorescence from nucleotide bound to DNAP. In addition, the size and shape of the ZMW prevents laser light from passing through the ZMW and limits the illumination to the bottom of the ZMW, which further increases the signal-to-noise ratio. As the four dNTPs are each labelled with a dif-

ferent fluorophore, each nucleotide can be identified from their unique fluorescence[60]. DNA base modification detection can also be achieved from analyzing DNAP kinetics, which consist of duration of fluorescence pulse, known as pulse width, and the duration between successive fluorescence pulses, referred to as interpulse duration [62]. To date, DNAP kinetics have been used to detect base modifications such as N6-methyladenine, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine [62] and DNA damage such as O6-methylguanine, 1-methyladenine, O4-methylthymine, 5-hydroxycytosine, 5hydroxyuracil, 5-hydroxymethyluyracil and thymine dimers [63].

SMRT platform capability was initially limited to continuous long read (CLR) generation with 10-15% error rate [60] instead of circular consensus sequence (CCS) generation with 0.1-1% error rate [64]. This was because there is an inherent trade-off between read length and read accuracy while DNAP processivity is held as a constant. The earlier generations of DNAP had insufficient processivity to sequence both the forward and reverse strand of a SMRTbell template multiple times. In contrast, the more recent generations of DNAP have sufficient processivity to sequence the forward and reverse strand of long SMRTbell templates (>10kb) multiple times such that both long and accurate reads are produced [64]. SMRT platform, hence, leveraged the improvements in DNAP processivity to first increase read length and subsequently improve read accuracy.

The PacBio RS instrument with the first generation of polymerase and chemistry (P1-C1) produced continuous long reads (CLR) with an average read length of 1,500 bp with 10-15% error rate []. In contrast, the most recent PacBio Revio instrument generates circular consensus sequence (CCS) reads with an average read length of 20,000 bp with 0.1-1% error rate []. In addition, the PacBio RS instrument used the first generation of SMRTcell with 150,000 ZMWs [] while the PacBio Revio instrument uses the latest SMRTcell with 25 million ZMWs, increasing the sequence throughput exponentially from 22 million bases to 90 billion bases per SMRTcell [] (Figure **??**). Compared to Illumina sequencing, CLR sequencing had a higher error rate and cost per-base, with only marginal increases in read length. In addition, the shortage of bioinformatics algorithms to process CLR reads with high error rate also slowed market adoption. The PacBio Revio instrument, however, can generate 30-fold CCS sequence coverage of the human genome under $1000. The sequence data from a single SMRTcell, therefore, can be used for not only *de novo* assembly [] but also haplotype-phased base modification[], SNP and indel, [] and structural variation detection [], enabling the most comprehensive characterisation of both genetic and epigenetic variation from a single human individual. We also expect the sequence throughput per SMRTcell to increase exponentially in the foreseeable future

with improvements in DNA processivity that increases CCS read length and advances in semiconductor fabrication technologies that doubles or triples the number of ZMWs per SMRTcell.

1.2.3 Long-read sequencing applications

In the beginning, long reads from ONT and PacBio SMRT platform did not have a competitive advantage compared to short reads from Illumina platform; long reads were only marginally longer than short reads and their higher error rate made accurate germline mutation detection more challenging. Long-read sequencing, most importantly, could not compete with short-read sequencing on sequencing cost.

1.2.3.1 *De novo* assembly

A substantial increase in read length from 1,500 bp to 10,000 bp with the introduction of XX chemistry for ONT and P5-C3 chemistry for PacBio Sequel I instrument reignited interest for new *de novo* assembly algorithm development, full-length transcript sequencing and accessing the inaccessible regions of the genome.

Genomes are peppered with repetitive sequences. These repetitive sequences, for example, account for more than 50% of the human genome[25]. The unique placement of a read in an assembly graph, therefore, requires read length to be longer than the repeat length such that unique sequences not found elsewhere in the genome flank the repetitive sequence in the read. Gaps and collapsed regions in genome assemblies, hence, often result from regions of the genome where the repeat length is longer than read length. There are, however, not many repeats except for segmental duplications[65], higher order repeats (HOR) in centromeres[66] and palindromic sequences in sex chromosomes that are longer than ONT and CLR reads [67].

A new generation of assembly algorithms based on de Brujin graph[68], string graph[69, 70] and OLC[71] were developed to leverage these long reads and enable end-to-end assembly of microbial genomes[72, 73] and large mammalian genomes[70, 71]. Complete hydatidiform mole (CHM) 1 BAC clones, for example, were selected for hierarchical shotgun sequencing to close existing gaps in the human reference genome [74]. At the time, contigs produced from these new assembly algorithms had unparalleled contiguity as measured by contig N50 []. In addition, misassemblies can be corrected, and contigs can be ordered and oriented into scaffolds using optical genome maps from Bionano Genomics [75]. Chromosome-length scaffold construction, more importantly, has become

routine through Hi-C scaffolding[76] and the ability to visualise[77] and manually inspect Hi-C contact matrix for assembly curation[78]. Trio-sequencing[79] and single-cell strand sequencing data[80] have also been used to also construct haplotype-resolved assemblies. These chromosome-length scaffolds, most importantly, are often comparable or better than existing reference genomes in both contiguity and completeness [81].

Ultra-long read library preparation from ONT and CCS library preparation from PacBio were two additional breakthroughs that transformed how *de novo* assembly is performed today. Ultra-long reads (>100kb) have been particularly useful for closing gaps[82] and for full-length sequencing of overlapping BAC clones for assembly of human chromosome Y centromere[83]. Human centromeres are enriched with AT-rich 171 bp tandem repeats called α-satellite DNA. Centromeric α-satellite DNA organises into HOR structures that are several megabases in length. Despite their crucial role in cell division, the organisation and structure of human centromeres were inaccessible to interrogation until the introduction of ultra-long reads. It is worth mentioning that centromeres in the b37 and hg38 reference genome are recorded as sequences and therefore do not provide a true representation of the underlying sequence [84].

CCS read length and accuracy have been leveraged to reduce computational complexity of all-to-all pairwise read alignments and shorten genome assembly time [85] and to distinguish recently diverged haplotypes and repeat copies such as segmental duplications [86, 87]. CCS reads are, routinely, used to produce haplotype-resolved chromosome-arm length contigs. It is worth mentioning that assembly algorithms often assume that the sample in question has a haploid genome. This assumption results in haplotype collapsed assemblies where the assembled haplotype is not present in the population [31]. The completion of telomere-to-telomere (T2T) CHM13 (T2T-CHM13) genome, including the short arms of five acrocentric chromosomes and centromeric satellite array, has been the culmination of years of effort to produce gapless and error-free assemblies [32]. These advancements allow us to construct high-quality reference genomes for a fraction of what it used to cost to build the human reference genome. The number of new plant and animal assemblies has burgeoned thanks to these developments [].

1.2.3.2 Full-length transcript sequencing

In contrast, to short-read sequencing that requires *de novo* assembly of RNA reads to acquire full-length transcripts, long-read sequencing can be used to obtain full-length transcript without assembly. Long-read sequencing has been used to successfully identify new isoforms in tissues and novel gene fusions in cancers []. Single-cell isoform-

sequencing has also been used to find new isoforms, to define the transcriptome atlas and to quantify the transcript in combination with single-cell RNA sequencing. In addition, these full-length transcripts have been successfully used for gene annotation of newly assembled genomes [].

1.2.3.3 Germline and somatic mutation detection

To date, ONT, CLR and CCS reads have been successfully used for germline SNP, small insertion and deletion[] and structural variation detection []. The lower base accuracy and higher per base sequencing cost has limited the use of ONT and CLR reads for SNP and indel detection. The longer read length, however, enabled access to regions of the genome inaccessible with short reads and early success in identification of pathogenic mutations in undiagnosed patients with rare diseases [].

Structural variation detection with short reads relies on either changes in sequence coverage for copy number variation (CNV) detection and identification of discordant read pairs with aberrant distance and orientation for breakpoint, translocation and inversion detection [88]. In contrast, long reads enable structural variation detection with nucleotide resolution through direct comparison of read and reference genome and is also more sensitive towards short tandem repeat (STR) expansions, short interspersed nuclear element (SINE) and long interspersed nuclear elements (LINE) insertion detection [89–91]. CHM1 CLR reads, for example, were also used to correct small misassembles in the reference genome and identify approximately 26,000 structural variations that were recalcitrant to detection using short reads [89]; the number of structural variations detected with long reads is at least double that detected with short reads. The number of structural variations is orders of magnitude smaller than the number of SNPs and indels, but structural variations alter greater number of bases and have a more pronounced impact on speciation and phenotype through gene regulation, duplication, translocation[92] and conformational changes in three-dimensional genome configuration[93**?**]. In addition, complex structural rearrangements such as chromothripsis[5, 94], chromoplexy[95] and templated insertions[96] are common oncogenic mechanisms. Repeat expansions and accompanied hypermethylation are common causes of neurological diseases[97]. The severity of Parkinson's disease, for example, is associated with repeat content and the size of the repeat expansion[]. Single-molecule sequencing is the only reliable technology for repeat expansion detection. Low genetic diagnosis rate of approximately 30% with short read sequencing and ability to detect haplotype phased genetic and epigenetic

variations with single molecule sequencing has renewed interest to detect causal and putative pathogenic mutations in patients with rare genetic disease[].

Despite the advantages that long-read sequencing technologies offers compared to short-read sequencing technologies for somatic structural rearrangement detection, the application of long-read sequencing technologies to somatic mutation detection has been limited to date. There has been a handful publications that interrogated somatic structural rearrangements in breast cancer cell lines with long reads []. Somatic mutation detection with long reads is at the stage where we are re-creating the capabilities provided by short-sequencing technology and is not at the stage where we are finding somatic mutations that cannot be detected with short-read sequencing technology.

Structural variation detection with short reads relies on either changes in sequence coverage for copy number variation (CNV) detection or identification of discordant read pairs with aberrant distance and orientation for breakpoint, translocation and inversion detection [88]. In contrast, long reads enable structural variation detection with nucleotide resolution through direct comparison of read and reference genome and are also more sensitive towards short tandem repeat (STR) expansions, short interspersed nuclear element (SINE) and long interspersed nuclear elements (LINE) insertion detection [89–91]. CHM1 CLR reads, for example, were also used to correct small misassemblies in the reference genome and identify approximately 26,000 structural variations that were recalcitrant to detection using short reads [89]; the number of structural variations detected with long reads is at least double that detected with short reads. The number of structural variations is orders of magnitude smaller than the number of SNPs and indels, but structural variations alter a greater number of bases and have a more pronounced impact on speciation and phenotype through gene regulation, duplication, translocation[92] and conformational changes in three-dimensional genome configuration[93**?**]. In addition, complex structural rearrangements such as chromothripsis[5, 94], chromoplexy[95] and templated insertions[96] are common oncogenic mechanisms. Repeat expansions and accompanied hypermethylation are common causes of neurological diseases[97]. The severity of Parkinson's disease, for example, is associated with repeat content and the size of the repeat expansion[]. Single-molecule sequencing is the only reliable technology for repeat expansion detection. Low genetic diagnosis rate of approximately 30% with short read sequencing and ability to detect haplotype phased genetic and epigenetic variations with single molecule sequencing has renewed interest to detect causal and putative pathogenic mutations in patients with rare genetic disease[].

1.3 The Darwin Tree of Life Project

1.4 Origins of Life

1.4.1 Prebiotic Earth

1.4.2 The Garden of Eden

1.4.3 RNA world

1.4.4 The emergence and evolution of life on Earth

The advent of high-throughput long-read sequencing[] and genome mapping technologies[], improvements in base accuracy of long reads [64] and development of algorithms that take advantage of the longer read length and long-range genomic interactions [76] have brought new enthusiasm to sequence and assemble high-quality reference genomes[].

The Darwin Tree of Life (DToL) project is an ambitious project that aspires to construct chromosome-length scaffolds for 70,000 eukaryotic species in Britain and Ireland []. In parallel, other international consortiums have initiated projects with similar aspirations for insects [], vertebrates [], invertebrates [] and all of life []. The DToL project, currently, uses CCS reads for contig generation, Hi-C reads to order and orient contigs, and a Hi-C contact matrix to manually inspect and correct chromosome-length scaffolds. The DToL project regularly updates their primary sequencing and mapping technologies and assembly, purging and scaffolding algorithms to reflect the advances in the field. At the time of writing, the DToL project has sequenced approximately 800 species, completed the assemblies of approximately 500 species, and made available to the public the raw data and reference genomes [].

1.5 Thesis objectives

The history of science is riddled with examples where theory, technology, and serendipitous discovery drive science. The advent of Illumina short reads and continued decrease in per-base sequencing cost have accelerated our understanding of human evolution and migration patterns[], identification of pathogenic mutations in patients with Mendelian diseases[], the analysis of driver mutation and transcriptomic landscape in thousands of cancer genomes[].

The inability to generate contiguous and complete reference genomes, however, with Illumina short reads[] and the prohibitively expensive cost of BAC clone library preparation and hierarchical shotgun sequencing have thwarted our efforts to understand genetic variation in non-model organisms[].

High-throughput and high-accuracy single-molecule sequencing technologies[64] overcome the limitations of the Illumina platform and propel us towards the third wave of genomic revolution where each individual will be able to have their complete and haplotype-phased genome sequence, where the construction of the most complex and repetitive genomes will be possible and where the reference genomes of all organisms will be available to the scientific community.

The DToL project, for example, has generated an extraordinary public resource that comprises CCS reads, linked reads, Hi-C reads, high-quality chromosome-length scaffolds, and associated gene annotations. Comparative genomics in linear and three-dimensional space and population genetic studies with the newly assembled reference genomes will undoubtedly enhance our understanding of the process of speciation and evolution. Here, we aspired to better understand the mutational process operational in each species.

To determine the germline and somatic mutational process across the Tree of Life, we considered the following:

1. Based on the similarities between the duplex[39] and CCS library sequencing [98] principles, we hypothesised that CCS reads might have sufficient base accuracy for ultra-rare somatic mutation and potentially single molecule somatic mutation detection.

2. CCS reads are reported to have a predicted accuracy above Q20, but their base accuracies have not been independently examined.

3. Somatic mutation detection algorithms need to distinguish somatic mutations from germline mutations in addition to sequencing, alignment and systematic bioinformatic errors.

4. Using samples with known ongoing somatic mutational processes and mutational signature analysis, we can demonstrate that CCS reads have sufficient or insufficient base accuracy for single molecule somatic mutation detection and determine the parameters that influence sensitivity and specificity.

5. If the sample in question has either high mutation rate or high mutation burden, the expected and the correct mutational spectrum will be observable from the validation and test data sets, respectively.

In short, we aimed to measure the CCS error rate, assess whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, develop a method to detect somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome and apply the method to understand germline and somatic mutational processes across the Tree of Life.

Chapter 2

Single molecule somatic mutation detection

2.1 Introduction

Based on my understanding of duplex sequencing methods [39, 40] and the recently developed nanorate sequencing protocol [24], a derivative of the duplex sequencing protocol and considering the similarities between two sequencing methods, I hypothesised that CCS reads might be as accurate or more accurate than duplex reads and that they can be used for single molecule somatic mutation detection.

Most sequences have been derived by priming on both strands; this allows more confidence than when only one strand could be used [99]

[Frederick Sanger]

The Sanger sequencing method can be described as one of the first-generation of sequencing methods and the original duplex sequencing method. The first iteration of the Sanger sequencing method required a single-stranded DNA template, a primer designed to bind to the start of the template DNA molecule, DNA polymerase to bind to the primer and initiate DNA synthesis, and free deoxyribonucleotides (dNTP) and dideoxynucleotides (ddNTP) to elongate and terminate DNA synthesis, respectively. The chain-termination experiment is repeated multiple times with four dideoxynucleotides (ddATP, ddGTP, ddCTP, ddTTP) to obtain DNA fragments of different sizes and DNA sequence is subsequently determined from reading the gel electrophoresis results from the four chain-termination experiments. Bi-directional Sanger sequencing can also be

performed to sequence both the forward and reverse strand of the template molecule and complementary base pairing between the two strands is leveraged to construct duplex reads with higher base accuracy [99]. To date, the Sanger sequencing method have been successfully used to obtain the 5,735 bp ΦX174 genome sequence [99] and reference genomes sequences of *D. melanogaster*, *C. elegans*, and *H. sapiens* [].

The current incarnation of the duplex sequencing method was developed for ultra-rare somatic mutation detection (<0.01% VAF) and to increase the limit of detection threshold beyond the technical limitations of the Illumina platform, in contrast to Sanger sequencing method that was used for genome assembly and germline mutation detection. The duplex library preparation protocol starts with the sonication and fragmentation of genomic DNA. Unique molecular identifier (UMI) consisting of 8 to 12 nucleotides and Illumina adapters are attached to double-stranded DNA molecules prior to their PCR amplification [39]. The duplex library is often diluted before PCR amplification to achieve optimal sampling and duplication per template molecule [40, 24]. PCR amplified library is sequenced using one of many Illumina sequencing instruments. Illumina reads are subsequently grouped according to their UMI and are classified as Watson or Crick strand depending on whether the sequence was derived from the P5 or P7 Illumina adapter, respectively.

A highly accurate duplex consensus sequence is, thereafter, generated leveraging the redundancies and complementary base pairing between the forward and reverse strand reads (Figure **??**). The higher sequence throughput of the modern Illumina instrument is critical in acquiring multiple reads (redundancies) from both strands of the template molecule and to identifying library errors introduced upstream of sequencing. DNAP, for example, might incorrectly replicate the template DNA molecule during PCR amplification, but the polymerase error will be present only in one copy or a subset of the copies. In addition, if both forward and reverse strand is sampled sufficiently, complementarity between the two strands can be used to demarcate bases with high accuracy from bases with low accuracy [39] and to estimate the base accuracy from the supporting bases and associated base quality scores [24]. Duplex reads, therefore, promises theoretical base accuracy of 1×10^{-9} (Q90), but in practice, duplex reads from the original protocol achieves base accuracy of 1×10^{-6} (Q60) [39].

In contrast, duplex reads from the nanorate library protocol achieves the promised Q90 base accuracy and single-molecule resolution somatic mutation detection [24]. To accomplish this, the nanorate library protocol identifies and addresses library errors upstream of PCR amplification to produce duplex libraries from error-free native DNA molecules. Blunt end restriction enzyme, for example, is used to fragment gDNA to

prevent enzymatic DNA misincorporation during end-repair and gap-filling. In addition, dideoxynucleotides are added to prevent single-strand displacement synthesis through nick translation, rendering DNA molecules that require this process unsuitable for library creation (Figure **??**). A highly accurate duplex read, thereafter, is constructed as described above.

CCS sequencing like duplex sequencing also takes advantage of the redundant sequencing and complementary base pairing between the forward and reverse strand to construct a highly accurate circular consensus sequence. The single-strand reads are referred to as subreads and an individual subread typically has 10-15% error rate [100]. CCS reads are reported to have an average read accuracy above Q20 [64], but their individual base accuracies have not been examined to date. I and others have hypothesised that PacBio circular consensus sequence (CCS) reads might be as accurate or more accurate than conventional duplex reads based on the similarities between the two protocols [] and the absence of PCR jackpot errors that occur in the earliest stage of PCR amplification. In addition, CCS reads have the added benefit of substantially longer read length (~10-20kb) that enables accurate placement of reads despite the presence of long repeats and allows more recently diverged repeats to be distinguished from each other in combination with the high base accuracy [].

CCS base quality score ranges from Q1 to nominal Q93, representing an error rate of 1 in 5 billion bases. If the BQ score estimates are correct, I imagined that single molecule somatic mutation detection will be possible across all human normal tissues, agnostic of clonality as the human genome accumulates ~17 somatic mutations per year per cell, equivalent to ~1 somatic mutation per human genome per 6 weeks [101]. In addition, in contrast to duplex sequencing methods where a matched normal sequencing is required to distinguish germline mutations from somatic mutations and where somatic mutation detection is limited to where restriction enzyme recognition site is available, CCS sequencing should enable genome-wide somatic mutation detection without a matched normal. If successful, haplotype-phased germline mutation (SNPs, indels and structural variations), epigenetic modifications and somatic mutation detection will be possible from bulk normal tissue CCS sequencing. This idea inspired us to assess the potential for single molecule somatic mutation detection using CCS reads where a single read alignment supports the mismatch between the read and the reference genome. Our understanding of somatic mutational processes across different tissue types was critical in selecting the samples to evaluate and demonstrate single molecule somatic mutation detection with CCS reads.

In short, I invert the premise that long reads are inaccurate and propose that CCS reads have the highest base accuracy among commercially available sequencing platforms. I assess the potential for single molecule somatic mutation detection using CCS reads, identify systematic errors with consensus sequence generation and base quality score estimation and propose potential solutions to address these issues. In addition, I present himut, a method that can call somatic mutations where a single read alignment supports the mismatch between the sample and the reference genome. I detail the rationale behind the mechanics of himut and report its sensitivity and specificity. I have designed himut with ease of use in mind, and himut requires a sorted BAM file with primary read alignments as the only input and returns a VCF file with somatic mutations as output. Himut is publicly available at https://github.com/sjin09/himut as a Python package under the MIT open license.

Single molecule somatic mutation candidates are generated from either a biological process or from a non-biological process such as library, sequencing, alignment, or systematic bioinformatics errors. If a single read supports the mismatch between the sample and the reference, somatic mutation is indistinguishable from errors. If, however, there is sufficient signal-to-noise ratio somatic mutation detection, mutational spectrum produced from the aggregate of somatic mutations should be consistent with the expected mutational signature for the sample.

I selected a set of samples (the BC-1 and HT-115 cell lines, as well as normal granulocytes from an 82-year-old female individual) as positive controls and a sample (cord blood granulocyte) with few somatic mutations as a negative control to determine the limit of detection, empirically calculate the CCS error rate and describe the CCS error profile. In contrast to a typical sample where multiple mutational processes might be active at any given time, single-cell clone expansion and sequencing studies have definitively identified APOBEC, POLE, clock-like mutational processes to be the dominant ongoing somatic mutational processes in BC-1, HT-115 and granulocytes, respectively [102, 101]. The mutational spectra from previous studies and the contribution of different mutational signatures to the mutational spectrum serve as truth sets to unbiasedly assess the accuracy of our somatic mutation detection algorithm and to experiment and evaluate the impact of different hard filters to sensitivity and specificity.

The APOBEC family of proteins is part of the innate immune response to viruses and retrotransposon. APOBEC enzymes acts upon single-stranded DNA and RNA as cytidine deaminase and catalyses cytosine to uracil deamination to deteriorate and initiate the degradation of the viral genome []. APOBEC mutational process inadvertently introduces

C>T (SBS2) and C>G/C>A (SBS13) mutations to the genome at TCN trinucleotides (Figure **??**) [] and localised hypermutations called kataegis, which are often observed at chromothriptic breakpoints []. APOBEC mutagenesis is, in fact, observed in more than 50% of human cancers and accounts for considerable proportion of the total mutational burden [].

DNA polymerase α (POLA), δ (POLD) and ε (POLE) cooperate to perform DNA replication. POLA is responsible for initiating DNA synthesis while POLD and POLE is responsible for bulk of DNA synthesis with high fidelity on the lagging and leading strand, respectively []. POLD and POLE enzymes both have intrinsic proofreading capabilities and their 3'-5' exonuclease activity removes 3'-terminal misincorporated nucleotide. Replicative DNA polymerases still introduce errors every 10^4–10^5 nucleotides, but the mismatch repair (MMR) machinery corrects these errors. Individuals with inherited germline mutations or acquired somatic mutations that inactivate the POLE exonuclease activity have elevated somatic mutation rate and predisposes them to polymerase proofreading-associated polyposis, endometrial and colorectal cancers []. C>A mutations at TCN trinucleotides (SBS10a), C>A/C>T mutations at TCN trinucleotides (SBS10b) T>G mutations at NTT trinucleotides (SBS28) (Figure **??**) characterise POLE mutagenesis [].

Clock-like mutational processes are mutational processes that introduces mutations at a constant rate throughout life and hence, number of mutations attributable to clock-like mutational processes is proportional to the age of the individual. Clock-like mutational process is sample and species dependent, but C>T (SBS1) mutations at NCG trinucleotide (Figure **??**) and cell division independent background mutational process (SBS5) (Figure **??**) [] are determined to be clock-like mutational processes in normal human samples. C>T mutations at CpG dinucleotide result from spontaneous deamination of 5-methylcytosine to thymine and the unrepaired T:G mismatch manifests as somatic mutations. The exact aetiology of SBS5 is unknown, but somatic mutagenesis study in post-mitotic tissues such as neurons and smooth muscle suggests that SBS5 might be a cell division independent process and that SBS5 might be a manifestation of multiple different mutational processes [].

2.2 Materials and Methods

2.2.1 CCS library preparation and sequencing

BC-1 and HT-115 cell lines were cultured in XX media containing XX and at XX in a humidified X environment. Umbilical blood (PD47269d) and peripheral blood sample of an 82-year-old female individual (PD48473b) were collected in 40-60mL lithium-heparin tubes and blood granulocytes were subsequently isolated using Lymphophorep. High molecular weight (HMW) DNA from BC-1 and HT-115 cell line and PD47269d and PD484873b blood granulocytes were extracted using Qiagen MagAttract HMW DNA extraction kit () and was sheared to 16-20kb DNA fragments using Megaruptor 3 system () with speed setting X. CCS sequencing libraries were constructed according to the 0.9.0 CCS library preparation protocol () and the libraries were sequenced using Sequel IIe instrument at the Wellcome Sanger Institute.

2.2.2 CCS read alignment and germline mutation detection

CCS reads with adapter sequences were identified with HiFiAdapterFilt [103] and were removed from downstream sequence analysis. CCS reads were aligned to the human reference genome (b37 and grch38) with minimap2 (version 2.24-r1155-dirty) with default parameters for CCS read alignment (-ax map-hifi –cs=short) [104] and primary alignments were selected, compressed, merged, and sorted with samtools (version 1.6) [105]. Germline SNPs and indels were detected with deepvariant (version 1.1.0) [106]. VCF files were compressed and indexed with tabix [107] and left aligned and normalised with bcftools (version 1.17-7-g097bda6) [108]

2.2.3 CCS empirical base quality calculation

To assess the potential for somatic mutation detection with CCS reads, we first assessed the accuracy of the BQ score estimate using CCS reads from cord blood granulocytes. The number of somatic mutations in cord blood granulocytes is limited to 40-50 somatic mutations per cell [109], and hence most SBS, excluding germline mutations, in cord blood granulocyte sample results from library, sequencing, alignment or bioinformatics error. The number of matches and mismatches were counted for each BQ score estimate to calculate the empirical BQ score. We considered reference allele and germline SNPs as matches and all other SBS as mismatches. Germline mutation detection using himut is described below. We excluded germline SNPs with genotype quality (GQ) score below

minimum GQ score of 20 and read depth above maximum depth threshold $4d+\sqrt{d}$, where d is the average read depth, from analysis. We, thereafter, calculated empirical BQ for each BQ score estimate (2.1):

$$\text{empirical BQ} = -10\log_{10}\left(\frac{\text{mismatch count}}{\text{match count}}\right) \tag{2.1}$$

To calculate the trinucleotide sequence context dependent CCS error rate, CCS reads from the cord blood sample were reconstructed, with the number of subreads for each CCS read set to 10 full-length subreads (the reasons are discussed in chapter 3). Cord blood CCS reads were subsequently processed as described above and below for read alignment and somatic mutation detection. To estimate the number of false positive mutations, the number of true positive somatic mutations were estimated from the number of callable bases and the cord blood somatic mutational process [101] and were subtracted from the number of trinucleotide sequence context normalised somatic mutation counts. The number of normalised false positive somatic mutation counts, and the number of callable trinucleotide bases were used to estimate the substitution and trinucleotide sequence context dependent CCS error rate.

2.2.4 Germline and somatic mutation detection

Germline and somatic mutations are both detected from bulk normal tissue leveraging CCS read length and base accuracy, characteristics unique to CCS reads and hard filters from previous publications [110, 111]. A BAM file with sorted primary read alignments is the only required input to obtain a VCF file with somatic mutations.

Upon initiation, read alignments are first randomly sampled from each target chromosome to compute the lower and upper bound read length and maximum read depth threshold $4d+\sqrt{d}$ where d is the average read depth. SBS candidates are collected from reads with average read accuracy, mapping quality score (MAPQ) and blast sequence identity greater than or equal to a predefined threshold. In addition, read length must be between the lower and upper bound read length to prevent somatic mutation detection from chimeric or fragmented reads. A naive Bayesian genotyper, thereafter, is applied to each SBS candidate to determine whether the data (D) only supports the variant as a germline mutation or whether the data support both a germline variant and a somatic mutation candidate simultaneously (2.2):

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)} \tag{2.2}$$

where $P(G)$ is the prior probability of observing the germline mutation genotype and D is the data that represents the pileup of read bases and corresponding sequencing error probabilities for each base at the substitution site. $P(D)$ is a constant across all the possible genotypes and is ignored. $P(G)$ is dependent on whether the genotype is heterozygous, heterozygous alternative (tri-allelic), homozygous alternative or homozygous reference allele with respect to the reference base (2.3):

$$P(G) = \begin{cases} \theta & \text{if } G = g_{\text{het}} \\ \frac{\theta}{2} & \text{if } G = g_{\text{hetalt}} \\ \theta^2 & \text{if } G = g_{\text{homalt}} \\ 1 - \frac{3\theta}{2} - \theta^2 & \text{if } G = g_{\text{homref}} \end{cases} \tag{2.3}$$

where θ is the expected germline SNP frequency and the default θ is set as 1×10^{-3}, the expected human germline SNP frequency.

$P(D|G)$ is the probability of observing the data given the genotype. Binomial likelihood is calculated for each genotype under the assumption that sequencing errors and read sampling is independent and identically distributed (2.4):

$$P(D|G) = \begin{cases} \frac{1}{2^n} \prod_i^n P(b_i|G) & \text{if } G = g_{\text{het}} \text{ or } g_{\text{hetalt}} \\ \prod_i^n P(b_i|G) & \text{if } G = g_{\text{homalt}} \text{ or } g_{\text{homref}} \end{cases} \tag{2.4}$$

where $P(b|G)$ is the probability of observing the base given the genotype and is defined as such 2.5

$$P(b_i|G) = P(b_i|A) = \begin{cases} 1 - \epsilon_i & \text{if } b_i \in A \\ \frac{\epsilon_i}{3} & \text{if } b_i \notin A \end{cases} \tag{2.5}$$

where b is CCS base covering the target locus, ϵ is the corresponding sequencing error probability and A is allele of the genotype. In practice, all calculations are performed in log scale. Phred scaled likelihood (PL) is calculated for the 10 possible genotypes (AA, CA, CC, CT, GA, GC, GG, GT, TA, TT) using the posterior probability of the genotype (2.6):

$$\text{PL} = -10 \log_{10} P(G|D) \tag{2.6}$$

and PL for each genotype is normalised using the lowest PL (2.7).

$$\text{normalised PL} = [\text{PL}_i, \text{PL}_{i+1}, \ldots, \text{PL}_{10}] - \text{PL}_i \tag{2.7}$$

where PL is assumed to be sorted from the smallest to the largest. The genotype with the lowest PL is selected as the germline genotype. Genotype quality (GQ) score of the selected germline genotype is the difference between the second lowest normalised PL and the lowest normalised PL. If the data only provides evidence for a germline mutation, the next SBS is then considered for somatic mutation detection. If the data support the presence of both a germline mutation and a somatic mutation candidate, a number of conservative hard filters are subsequently applied to distinguish somatic mutations from errors:

1. If the germline mutation is a heterozygous, heterozygous alternative or homozygous alternative allele, somatic mutation candidate is excluded from the downstream analysis as somatic reversions are not considered. Somatic mutation detection, hence, is restricted to locus with homozygous reference allele to prevent the misclassification of heterozygous mutation as a somatic mutation.

2. The GQ score for the homozygous reference allele needs to be above the minimum GQ score threshold.

3. The BQ score of the somatic mutation candidate needs to be above the minimum BQ score threshold.

4. Indels must be absent from the SBS locus.

5. The read depth of the target locus needs to be below the maximum depth threshold.

6. The reference allele count and the alternative allele count need to be above the minimum reference allele and alternative allele count. This condition is not required if the sample has sufficient sequence coverage as the GQ score is positively correlated with sequence coverage.

7. CCS reads with adapter sequences might still be present in the BAM file and therefore the somatic mutation candidate might result from incomplete adapter trimming. Candidates located in close proximity to start and ends of reads are filtered as specified with the --min_trim parameter.

8. The number of mismatches adjacent to the candidate needs to be below the --max_mismatch_count threshold as an alignment error can be mistaken as a somatic

mutation. The size of the window is specified with the --mismatch_window parameter.

A VCF file with common SNPs (>1% major allele frequencies) and a Panel of Normal (PoN) VCF file can also be optionally provided to exclude somatic mutation candidates potentially resulting from DNA contamination and systematic bioinformatics error, respectively. In addition, a VCF file with haplotype-phased hetSNPs can be provided to limit somatic mutation detection from haplotype phased CCS reads. Here, himut with default parameters (--min_qv 30 --min_sequence_identity 0.99 --min_gq 20 --min_bq 93 --min_trim 0.01 --min_ref_count 3 --min_alt_count 1 --min_hap_count 3 --mismatch_window 20 --max_mismatch_count 0) were used for the identification of unphased and haplotype phased somatic mutation. As sex chromosomes are enriched for misassembled regions and repetitive sequences [], somatic mutation detection was restricted to the autosomes. To process BAM, FASTA/Q and VCF files, himut internally uses pysam [112], pyfastx [113] and cyvcf2 [114], respectively. In addition, multiprocessing Python package was used to enable parallel processing.

2.2.5 Panel of Normal construction

We created a PoN VCF file from 11 normal individuals with publicly available CCS dataset (Table X) to reduce the number of false positives arising from systematic bioinformatics errors. We ran himut with relaxed parameters (--min_mapq 30 –-min_trim 0 –-min_sequence_identity = 0.8 –-min_hq_base_proportion 0.3 –-min_alignment_proportion 0.5 –-min_bq = 20) to maximise the number of mutations called from these samples. The number of samples in the PoN VCF is currently limited to the number of publicly available CCS dataset. As the number of CCS sequenced samples increases, in the future the power to distinguish somatic mutations from artefacts will also increase.

2.2.6 Germline mutation haplotype phasing

A haplotype is defined as a group of genetic variations that are inherited together from a single parent. We treat haplotype phasing as a graph algorithm problem where each hetSNP is a node in a graph and there is an edge between a pair of haplotype consistent hetSNPs. A single CCS read spans multiple heterozygous SNPs and evidence from multiple CCS reads can determine whether a pair of hetSNP is haplotype consistent ($p < 0.0001$ one-sided binomial test) (Figure**??**). If a pair of hetSNP is haplotype consistent, a pair of hetSNP

exists in cis configuration or trans configuration (Figure**??**). A haplotype inconsistent pair of hetSNP results from non-biological sources. Haplotype consistency is measured between all possible hetSNP pairs and hetSNP that is haplotype consistent with at least 20% of its possible pairs is connected through breadth-first search algorithm to construct contiguous haplotype blocks. Himut accepts as input VCF file from deepvariant and returns a VCF file with haplotype-phased hetSNPs.

2.2.7 Haplotype-phased somatic mutation detection

CCS reads are assigned to a haplotype block to enable haplotype-phased somatic mutation detection. To be allocated to a haplotype block, a CCS read must be within a haplotype block (and not between two haplotype blocks) and have haplotype identical to the consensus haplotype as defined in the haplotype block. In essence, somatic mutations are not phased through adjacent hetSNPs and instead phased CCS reads are used for somatic mutation detection. In addition, haplotype counts from the wild type CCS reads without the somatic mutation need to be above minimum haplotype count threshold to prevent misclassification of hetSNPs as somatic mutations.

2.2.8 Somatic mutation count normalisation and mutation burden calculation

To determine the number of somatic mutations called per cell, the number of somatic mutations needs to be normalised with the number of callable CCS bases and reference bases. We use the same conditions as somatic mutation detection to count the number of callable CCS bases and callable reference bases and derive the trinucleotide sequence context frequency for the callable genomic bases and the callable CCS bases.

We, thereafter, normalise the somatic mutation counts with the ratio of reference trinucleotide sequence context frequency to the callable reference trinucleotide sequence context ferquency and the ratio of callable reference trinucleotide context frequency to callable CCS trinucleotide sequence context frequency.

2.2.9 CCS base quality recalibration

Subreads are aligned to CCS reads from the same ZMW using ACTC [] to determine the orientation of subreads and identify fragmented or concatenated subreads. DeepConsensus accepts the BAM file with subreads aligned to CCS read as input, polishes CCS reads

and recalibrates the BQ score. CCS reads and subreads from the same ZMW were used to construct partial order alignment using abPOA [115] and the resulting alignment was processed to identify CCS bases where there is unanimous support from the subreads. CCS bases with unanimous support were assigned Q93 BQ score while all other bases without unanimous support were assigned Q0 BQ score.

2.3 Results

2.3.1 CCS library errors and sequencing errors

CCS reads have been successfully used for construction of highly contiguous and complete de novo assemblies [] and germline mutation detection []. 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 the base accuracy must be higher than the human genome somatic mutation rate (1-2 mutations per 1-4 weeks per cell), equivalent to approximately ~Q90 to distinguish sequencing errors from single molecule somatic mutations. In addition, library, sequencing and systematic errors and genomic DNA contamination are common sources of false positive somatic mutations.

We generated 30-fold CCS sequence coverage from BC-1, HT-115 and blood granulocytes from an 82-year-old female individual (PD48473b) and 70-fold CCS sequence coverage from cord blood granulocyte (PD47269d) with an average read length between 16 and 20kb (Table 2.1) to achieve the following objectives: 1), assess the potential for single molecule somatic mutation detection with 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, 4) develop a method for somatic mutation using CCS reads and 5) assess the sensitivity and specificity of our method.

We, first, examined the library preparation and circular consensus sequence construction process to minimise the number of library and sequencing errors. HMW DNA for CCS library preparation is often prepared through Qiagen Magattract or Circulomics HMW DNA extraction kit and HMW DNA is sheared to the appropriate size using a Megaruptor instrument. A hairpin adapter is attached to both ends of the double-stranded DNA molecule to create a topologically circular template. DNA nuclease is subsequently used to digest DNA molecules (e.g, failed ligation products) not suitable for sequencing. Primer

Table 2.1 Experimental Data

	BC-1	HT-115	PD47269d	PD4873b
Genomic DNA source	Cell line		Blood granulocyte	
Age (years)	-	-	0	82
CCS read count	5,962,252	5,933,281	12,156,251	4,949,180
Mean length ± std (bp)	18,571 ±	17,038 ±	16,523 ± 3,752	18,263 ± 1,753
Q93 bases (%)	51.4	55.5	57.6	51.7
Sequence coverage	36.9	33.7	67.0	30.1
Mutational process	APOBEC	POLE	Clock-like	
Mutational signature	SBS2	SBS10a, SBS10b and SBS28	SBS1 and SBS5	
Mutation burden per cell	~2,000 - 22,000	~8,000 - 11,000	~40 - 50	~1400 - 1500

with poly-A tail, thereafter, is annealed to the hairpin adapter sequence. BluePippin based size selection may additionally be performed to prepare size-selected libraries to maximize sequence throughput per SMRTcell.

A DNA damage repair enzyme cocktail (unpublished) is used to repair DNA damage (nicks, abasic sites, thymidine dimers, blocked 3'-ends, oxidised guanine and pyrimidines and deaminated cytosines) introduced during library preparation (personal communication). In addition, end-repair and A-tailing is performed to remove protruding ends and to enable adapter ligation, respectively. Defective DNA damage repair or unrepaired DNA damage manifest as library errors and can be misclassified as a somatic mutation. The precise identity of DNA damage repair enzymes in the cocktail are unknown. We, however, can make informed assumptions about their function and their impact on downstream sequence analysis, and highlight the DNA damage repair process that is most likely to introduce library errors. Nanoseq protocol, for example, pinpoints end-repair and nick translation processes to be the primary sources of library errors. Strand-displacement synthesis during nick translation, for example, can introduce kilobases of sequences using the complementary strand as a template (Figure **??**) [].

CCS libraries are loaded on the SMRTcell and template DNA molecules diffuse into one of the ZMWs. A productive ZMW is defined as a ZMW with a single template molecule, from which a sufficient number of subreads are sequenced to construct a consensus sequence with at least Q20 average read accuracy. DNAP at the bottom of the ZMW binds to the DNA primer and initiates rolling circle amplification through strand-displacement synthesis. DNAP incorporates fluorescently labelled nucleotides, fluorescence emitted during DNA incorporation is measured and fluorophore is cleaved off upon successful incorporation. The wavelength of the fluorescence, length of the fluorescence, and dura-

tion between the successive pulses of fluorescence is used to determine the identity of the base and chemical modifications to the base.

The DNAP from the latest library protocol has sufficient processivity to generate an average of 10-12 full-length subreads on average for template molecules with read-of-insert length between 16kb and 20kb. The single-strand readouts of the forward and reverse strand of the template molecule are referred to as subreads. The first subread and the last subread are often partial readouts of the template molecule because of internal priming and sequencing termination respectively, while the subreads from the second to the second-to-last subreads are full-length readouts of the template molecule (Figure **??**). Assuming perfect detection of adapter sequences, odd-numbered subreads and even-numbered subreads are assumed to have the same sequence orientation as DNAP is agnostic to strand orientation. A draft sequence is constructed from multiple sequence alignment of subreads and is polished based on the realignment of subreads to the draft sequence. A dinucleotide sequence context Hidden Markov Model (HMM) is used to infer the base accuracy and DNA sequence from the observed subread bases (personal communication). A highly accurate consensus sequence can be constructed as sequencing errors are assumed to be randomly introduced without sequence context bias and are independent of one another. In addition, non-complementary base pairing between the forward and reverse strand indicates the presence of either a library or a sequencing error and resulting CCS is assigned a low BQ score.

PacBio circular consensus sequence algorithm (pbccs) calculates the median subread length and uses subreads with read length above 50% of median subread length and below 200% of median subread length for CCS generation for CCS generation. If adapter sequences are incorrectly detected within the subread or if adapter sequences are not detected where present, full-length subreads can be fragmented into multiple partial subreads and multiple subreads can be concatenated into a single subread, respectively. Unfortunately, read length based hard filters cannot identify all cases where adapter sequence detection has failed.

To identify potential errors introduced during CCS library preparation and sequencing, CCS and subreads from the same ZMW were analysed together and sequence quality control was performed (Methods). We observed that X% of ZMWs have fragmented and/or concatenated subreads (Figure **??**). We hypothesise that CCS reads with read length deviating from mean CCS read length are the result of failed adapter sequence detection and exclude these CCS reads from somatic mutation detection (Method). In

addition, we also noticed higher than expected adenine and thymine proportion at the end of CCS reads resulting from incomplete adapter sequence trimming (Figure **??**)

CCS reads have an average read accuracy of at least Q20 and individual BQ score ranges from Q1 to nominal Q93, corresponding to 0.5×10^{-9} error rate (Figure **??**). To our knowledge, the accuracy of CCS BQ has not been examined to date. CCS read accuracy and BQ score is dependent on the number of subreads per CCS read (Figure **??**) and concordance between the subread bases and the CCS base. We confirm that the number of substitutions and indels is negatively correlated with CCS read accuracy and the number of subreads per CCS read as reported in a previous publication (Figure **??**). The accuracy of the BQ score, hence, is expected to increase with the number of supporting subread bases. We, however, observed that the accuracy of the CCS BQ score decreases with increase in the number of subreads and that increase in the number of subreads per CCS read results in not diminishing returns, but negative returns to CCS base accuracy (discussed later in Chapter 3). To determine whether CCS bases have sufficient base accuracy for single molecule somatic mutation detection, we measured the empirical BQ score using cord blood CCS reads and (Methods) and ascertained that CCS bases have sufficient accuracy for rare somatic mutation detection where a sample has a high mutation burden or a high somatic mutation rate (Figure **??**). Using positive control samples, we identified additional CCS read characteristics that influences somatic mutation detection sensitivity and specificity.

2.3.2 Germline mutation and somatic mutation detection

Somatic mutagenesis is a continuous process throughout life. Bulk normal tissue has germline mutations that are inherited from parents, mosaic mutations that occurred during embryonic development and newly acquired somatic mutations from ongoing mutational processes. In addition, cells with driver mutations can outcompete neighbouring cells and undergo clonal expansion. Paired tumour-normal sequencing is often performed to distinguish germline mutations from somatic mutations in a tumour sample. Here, we present how we distinguish errors and germline mutations from somatic mutations in bulk normal tissue, leveraging CCS read length and base accuracy.

We, first, compared germline SNPs detected from both himut and deepvariant to assess whether our algorithm can accurately call genetic variations (Table). The number of SNPs and transition to transversion (TiTv) ratio is within the expected range, demonstrating that himut can also function as a standalone variant caller. We believe that algorithmic

differences account for disparities in number of SNPs called with himut and deepvariant, which is a deep learning based variant caller that uses read pileup images for germline mutation detection while himut uses an analytical approach similar to GATK for germline mutation detection.

To distinguish germline mutations from somatic mutations, himut detects and classifies germline mutations as heterozygous, heterozygous alternative, homozygous alternative, or homozygous reference allele (Method, Figure **??**). Somatic mutation candidates are collected from CCS reads meeting the defined read-level prerequisites and candidates are categorised according to their base-level conditions (Figure **??**). Somatic mutation detection is also restricted to homozygous reference allele sites as somatic reversions might be the result of DNA contamination. To calculate the mutation burden of the sample, himut identifies the number of callable bases using the same conditions as somatic mutation detection and normalises the somatic mutation count based on the number of callable CCS bases and reference bases (Method). A VCF file with haplotype phased heterozygous SNPs (hetSNPs), a VCF file with common SNPs and a PoN VCF file can also be optionally provided to call haplotype phased somatic mutations, to exclude false positive mutations resulting from DNA contamination and discard false positive mutations arising from systematic errors, respectively.

CCS read length and base accuracy can also be leveraged to phase hetSNPs and construct contiguous haplotype blocks, which enables haplotype phasing of CCS reads and haplotype-phased somatic mutation detection (Method). Read-backed phasing with Illumina reads uses adjacent hetSNPs to phase approximately ~30% of detected somatic mutations []. In contrast, haplotype phased somatic mutation detection with CCS reads uses all hetSNPs that CCS read spans and phases approximately ~ 70% of somatic mutations (Figure **??**). In addition, haplotype phased somatic mutation detection has three advantages: 1) CCS reads derived from DNA contamination often do not possess the same haplotype as the sample. If CCS read do not share the consensus haplotype, CCS read is excluded from somatic mutation detection (Figure, 2**??**) If two haplotypes are unevenly sampled, hetSNP can be misclassified as somatic mutations in low coverage samples. Restricting somatic mutation detection to haplotype phased regions limits somatic mutation detection to regions where both haplotypes have been adequately sampled. (Figure **??**), 3) CCS read with the same somatic mutations should share the haplotype and somatic mutations should not be present on both haplotypes (Figure **??**). Haplotype phased somatic mutation detection is especially helpful for samples with high heterozygosity.

2.3.3 Somatic mutation detection sensitivity and specificity

We called and benchmarked haplotype phased and unphased somatic mutations from the three positive controls with different mutational burdens and distinct mutational processes. Our unique benchmarking approach leverages the fact that a single somatic mutational process is active in each sample and that somatic mutation candidates are derived from either errors or newly acquired somatic mutations. We cannot be certain whether individual somatic mutations are derived from a biological process or a non-biological process, but the mutational spectrum produced from the aggregate somatic mutations should be consistent with the expected mutational signature, if there is sufficient signal-to-noise ratio for somatic mutation detection. In addition, our approach is not biased towards Illumina callable regions of the genome unlike the Genome in a Bottle (GIAB) benchmarks [] as our somatic mutation detection method is agnostic to reference position.

We calculated mutation burden from BC-1, HT-115 and PD48473b samples to be X, X, X, respectively, consistent with previous estimates []. In addition, high cosine similarity between the expected mutational signatures and mutational spectrum from our positive control samples demonstrate that PacBio CCS bases have sufficient base accuracy for rare somatic mutation detection where samples have a high mutation burden or high somatic mutation rate (Method). Moreover, we can determine the number of true positive mutations and false positive mutations from the called somatic mutations and the number of true negative mutations and false negative mutations from the filtered somatic mutations through mutational signature analysis. We can subsequently use these estimates to calculate the sensitivity, specificity, specificity, and F1-score for each of our samples (Method, Table). We also selected appropriate hard filter thresholds based on receiver operating characteristic (ROC) curves generated under a range of hard filter conditions (Figure **??**) and determined hard filters with the greatest impact on sensitivity based on odds ratio calculated in the absence and presence of the hard filter in question. The minimum BQ and GQ scores were crucial for somatic mutation detection while other filters had a marginally positive impact on somatic mutation sensitivity. We would like to also highlight that somatic mutation detection sensitivity and specificity increased when grch38 was used as a reference genome, reflecting better representation of genetic polymorphisms with improvements in assembly quality (Table). We, unfortunately, could not compare himut with other methods as himut is the first somatic SBS detection method with CCS reads and as somatic mutation detection below 0.1% VAF has not been technically feasible with Illumina reads.

2.3.4 CCS errors, error rate calculation and base quality score recalibration

The mutation burden in the cord blood sample is the lowest, with only 40-50 somatic mutations per cell []. CCS bases, unfortunately, do not have sufficient signal-to-noise ratio to enable somatic mutation detection in the cord blood sample with high confidence. Mutational spectrum from the cord blood sample, which we refer to as the CCS error profile, is dissimilar to the expected mutational signature as the number of false positive mutations exceeds the number of true positive mutations (Figure **??**). CCS error profile occurs in multiple samples, suggesting that the error process is systematic in nature (Figure **??**). Using the number of false positive mutations and the callable number of bases, we calculated the CCS error rate to range from Q60 to Q90 depending on the substitution and the trinucleotide sequence context (Method, Figure **??**).

Library, sequencing, and software error upstream of somatic mutation detection are potential sources of false positive mutations. We triangulated software error as the origin of the CCS error profile through somatic mutation detection using uncapped BQ scores, deepConsensus polished CCS reads [] and CCS reads with recalibrated BQ scores (Method, Figure **??**).

CCS BQ score ranges from Q1 to Q93 and BQ scores are encoded with the ASCII character encoding format. BQ score is capped at Q93 because ASCII characters cannot support Phred-scaled quality values (QV) above 93. Inability to detect somatic mutations accurately with uncapped BQ scores demonstrates that there is a persistent problem with BQ score estimation (Figure **??**).

DeepConsensus calculates BQ score based on alignment of subreads to the CCS read from the same ZMW and BQ score of deepConsensus polished CCS reads ranges from Q1 to Q50 (Figure **??**), which we think is too conservative considering the empirical BQ score estimation from the cord blood sample. We also observed that somatic mutation detection with polished Q50 CCS bases did not generate the expected mutational spectrum while that with polished CCS bases with BQ score above Q30 generated the expected mutational spectrum, suggesting that once again BQ score is not accurately estimated.

To assess potential for single molecule somatic mutation detection with CCS reads, we performed partial order alignment between CCS read and subreads from the same ZMW and identified bases where there is unanimous support for the CCS base from the subreads (Method). Somatic mutation detection with CCS bases with unanimous support from subreads generates the expected mutational spectrum from the cord blood

sample, suggesting that software error and not sequencing error is the source of false positive mutations. We hypothesise that the PacBio consensus sequence construction and polishing algorithm consider somatic mutations as errors and as a result have incorrect sequencing error priors and BQ score estimates.

2.4 Conclusion

Here, I assess whether CCS reads are as accurate as duplex reads and demonstrate that a subset of CCS bases has sufficient base accuracy to enable single molecule somatic mutation detection using samples with single ongoing somatic mutational process. Himut takes as input a sorted BAM file with primary read alignments from bulk normal tissue, leverages CCS read length and base accuracy to distinguish somatic mutations from errors and germline mutations and returns a VCF file with somatic mutations. Mutational spectrum produced from aggregate of somatic mutations is concordant with the expected mutational signature from each positive control sample, showing that single molecule somatic mutation detection is indeed possible with CCS reads.

Using a cord blood sample with few somatic mutations, I examined the nature of residual false positive substitutions and associated CCS error profile that is shared across all samples. I empirically estimated that CCS Q93 base accuracy ranges from Q60 to Q90 depending on the substitution and trinucleotide sequence context, which is hundred thousand-fold to a billion-fold more accurate than Illumina bases and what enables somatic mutation detection with high confidence.

I conclude that false positive mutations are in fact derived from a combination of software errors. I show the persistence of inaccurate BQ score estimates using a modified pbccs that returns uncapped base quality scores, deepConsensus polished CCS reads and BQ score recalibration from partial order alignment between subreads and CCS reads from the same ZMW. I unexpectedly found that BQ score estimate becomes more inaccurate as the number of supporting subreads per CCS reads increases in contrast to the expected behaviour of the software (discussed and demonstrated in Chapter 3). In addition, I observe that false positive substitutions are enriched trinucleotide sequence contexts where the 5' base or the 3' base is identical to the substitution error. I hypothesize that inappropriate sequencing priors and underestimation of somatic mutations as potential sources of error in accurate BQ score estimation, and the use of trinucleotide sequence context HMM instead of dinucleotide sequence context HMM might ameliorate some of the issues. I, most importantly, show that subreads have sufficient base accuracy to

generate CCS bases with ~Q90 base accuracy at all trinucleotide sequence contexts, if there is enough supporting subreads per CCS read.

2.5 Discussion

I conjecture that issue with CCS BQ score estimation will be properly addressed and that majority of CCS bases will have ~Q90 base accuracy in the imminent future. I, here, discuss the ramifications and potential applications following this development.

2.5.1 Somatic mutation detection

To date, CCS reads have been successfully used for construction of chromosome-length scaffolds of microbial and eukaryotic genomes [], used for germline SNP, indel and structural variation detection [], and have improved the genetic diagnosis rate of rare diseases []. The applications of CCS read for somatic mutation detection, however, have been limited and there has only been a handful of publications studying the complex structural rearrangements in cancers using CCS reads []. Here, I focused on single molecule somatic SBS detection with the intention to identify and analyse somatic mutational processes across the Tree of Life (discussed in Chapter 3) while others focused on improving the sensitivity and specificity of structural variations that could already be detected with Illumina reads []. Himut still cannot distinguish whether an individual SBS is an error or a somatic mutation, but posterior probability can be calculated to determine the probability that the substitution is derived from a biological process or a non-biological process (**??**).

This approach previously was used to determine SBS16 mutational signature, a signature associated with alcohol consumption, as the main source of somatic mutations in CTNNB1 gene in hepatocellular carcinoma []. Despite this problem, himut will still enable researchers to rapidly screen for mutational signatures from bulk normal tissue without arduous experiments such as LCM or single-cell clone expansion sequencing, identification of environmental mutagenesis such as exposure to aristolochic acids[] across different locations and populations, lineage trace embryonic and tumour development through accurate detection of mosaic and somatic mutations, respectively. In addition, the ability to calculate the mutation burden in normal samples and thereby the age of the samples also raises the interesting question with regards to how to protect individual's privacy when SMRT platform becomes the primary sequencing method.

Himut currently does not consider matched tumour-normal sequencing for somatic mutation detection, but this would be the natural next step as the number of matched tumour-normal samples sequenced with the SMRT platform is expected to increase with the introduction of the Revio instrument. In the future, when error-free native DNA CCS library preparation is possible and when CCS BQ scores are correctly calibrated, HMW DNA extraction, input requirements for CCS library preparation and sequence coverage of the sample becomes the limiting factor to identifying and studying somatic mutagenesis across all tissues and all species.

In the interim, I believe that a wider range of somatic mutation detection will be possible with the benchmarking approach I have established where a sample with a known double base substitution and indel somatic mutational process is sequenced and used to fine-tune the pbccs algorithm and improve himut sensitivity and specificity. UV light, for example, induces the photoexcitation and dimerisation of adjacent pyrimidines into cyclobutane pyrimidine dimer (CPD) and 6-4 photoproduct. Although the exact mechanism that converts DNA damage to DNA mutation is unknown, CPD deamination has been suggested as one of the mechanisms generating C>T mutations (SBS7abc) and CC>TT mutations (DBS1) []. Cisplatin, a commonly used chemotherapy drug, forms inter-strand DNA crosslinks to prevent DNA replication, which induces cell cycle arrest and apoptosis. Cisplatin produces a unique mutational signature where a single T insertion is introduced downstream of GG dinucleotides [], which is attributed to nucleotide excision repair of 1-3d(GpXpG) intra-strand cisplatin adducts []

2.5.2 Strand-specific somatic mutation detection

Somatic mutation is a three-step process: 1) DNA damage or modification from exogenous or endogenous sources, 2) failure to detect and repair the DNA damage correctly,and 3) fixation and persistence of DNA mutation in daughter cells. Mutational signature is a mathematical abstraction of these three inter-dependent processes (**??**) and describes the probability that a given somatic mutational processes will introduce a mutation at a specific sequence context.

The unique capability of SMRT platform to both generate single-strand consensus sequences (SSCS) and CCS reads, along with the ability to detect epigenetic modifications [**? ?**] and somatic mutations at a single-molecule resolution using CCS reads, presents an exciting opportunity to dissect each mutational signature in more detail.

DNA damage and repair process associated with SBS1 mutational signature, for example, is amenable to further qualitative and quantitative examination through CCS sequencing. The spontaneous deamination of 5mC to thymine results in a TG:GC mismatch, which results in C>T mutations at CG dinucleotides if left unrepaired by the mismatch repair (MMR) pathway. SSCS reads enable the detection of TG:GC mismatches and genome-wide mapping of DNA damage. CCS reads allow the simultaneous detection of 5mC base modification and C>T somatic mutations at single-molecule resolution (Figure **??**). If successful, we will be able to measure the *in vivo* deamination rate and compare it against the *in vitro* deamination rate of $5.8x10^{-13}$ per 5mC per second at 37°C [], measure DNA mismatch repair efficiency and fidelity under mutant and wild-type conditions, and estimate the nonlinear contribution of DNA damage, repair and mutation fixation process to the SBS1 mutational signature. MutSα deficiency, for example, elevates the number of C>T somatic mutations, demonstrating the critical role of MutSα in recognising the TG:GC mismatch and initiating MMR. Cross-examination of both SSCS and CCS reads and associated DNAP kinetics might also allow us to better understand the APOBEC mutagenesis (SBS2) resulting from cytosine to uracil deamination.

2.5.3 Gene conversion and crossover detection

I, here, also hypothesise that CCS read length and base accuracy can be leveraged for genome-wide gene conversion and crossover detection resulting from double-strand break (DSB) repair and characterise the differences between meiotic and mitotic recombination products.

Chapter 3

Germline and somatic mutational processes across eukaryotic species in the Darwin Tree of Life project

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 cer-

tain 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 detected 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

Conclusions

4.1 Summary of findings

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.

4.2 Limitations

4.3 Future directions

4.4 Concluding remarks

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

[1] Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, April 2009.

[2] Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer*, 20(10):555–572, October 2020.

[3] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[4] Sam Behjati, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C Wedge, Asif U Tamuri, Inigo Martincorena, Mia Petljak, Ludmil B Alexandrov, Gunes Gundem, Patrick S Tarpey, Sophie Roerink, Joyce Blokker, Mark Maddison, Laura Mudie, Ben Robinson, Serena Nik-Zainal, Peter Campbell, Nick Goldman, Marc van de Wetering, Edwin Cuppen, Hans Clevers, and Michael R Stratton. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, September 2014.

[5] Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, Stuart McLaren, Meng-Lay Lin, David J McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P Butler, Jon W Teague,

Michael A Quail, John Burton, Harold Swerdlow, Nigel P Carter, Laura A Morsberger, Christine Iacobuzio-Donahue, George A Follows, Anthony R Green, Adrienne M Flanagan, Michael R Stratton, P Andrew Futreal, and Peter J Campbell. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, 144(1):27–40, January 2011.

[6] Peter Armitage and Richard Doll. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br. J. Cancer*, 8(1):1–12, March 1954.

[7] A G Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, 68(4):820–823, April 1971.

[8] Glenn M Marshall, Daniel R Carter, Belamy B Cheung, Tao Liu, Marion K Mateos, Justin G Meyerowitz, and William A Weiss. The prenatal origins of cancer. *Nat. Rev. Cancer*, 14(4):277–289, April 2014.

[9] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, September 2013.

[10] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020.

[11] D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000.

[12] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

[13] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, 3(1):246–259, January 2013.

[14] Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, Marcin Imielinski, PCAWG Structural Variation Working Group, Joachim Weischenfeldt, Rameen Beroukhim, Peter J Campbell, and PCAWG Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, February 2020.

[15] Christopher D Steele, Ammal Abbasi, S M Ashiqul Islam, Amy L Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, Dolapo Ajayi, Annelien Verfaillie, Pawan Dhami, Alex McLatchie, Matt Lechner, Nicholas Light, Adam Shlien, David Malkin, Andrew Feber, Paula Proszek, Tom Lesluyes, Fredrik Mertens, Adrienne M Flanagan, Maxime Tarabichi, Peter Van Loo, Ludmil B Alexandrov, and Nischalan Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, June 2022.

[16] Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, S M Ashiqul Islam, Nuria Lopez-Bigas, Leszek J Klimczak, John R McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A Wheeler, Ville Mustonen, PCAWG Mutational Signatures Working Group, Gad Getz, Steven G Rozen, Michael R Stratton, and PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, February 2020.

[17] Andrea Degasperi, Xueqing Zou, Tauanne Dias Amarante, Andrea Martinez-Martinez, Gene Ching Chiek Koh, João M L Dias, Laura Heskin, Lucia Chmelova, Giuseppe Rinaldi, Valerie Ya Wen Wang, Arjun S Nanda, Aaron Bernstein, Sophie E Momen, Jamie Young, Daniel Perez-Gil, Yasin Memari, Cherif Badja, Scott Shooter, Jan Czarnecki, Matthew A Brown, Helen R Davies, Genomics England Research Consortium, and Serena Nik-Zainal. Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science*, 376(6591), April 2022.

[18] Oriol Pich, Ferran Muiños, Martijn Paul Lolkema, Neeltje Steeghs, Abel Gonzalez-Perez, and Nuria Lopez-Bigas. The mutational footprints of cancer therapies. *Nat. Genet.*, 51(12):1732–1740, December 2019.

[19] Sarah J Aitken, Craig J Anderson, Frances Connor, Oriol Pich, Vasavi Sundaram, Christine Feig, Tim F Rayner, Margus Lukk, Stuart Aitken, Juliet Luft, Elissavet Kentepozidou, Claudia Arnedo-Pac, Sjoerd V Beentjes, Susan E Davies, Ruben M Drews, Ailith Ewing, Vera B Kaiser, Ava Khamseh, Erika López-Arribillaga, Aisling M Redmond, Javier Santoyo-Lopez, Inés Sentís, Lana Talmane, Andrew D Yates, Liver Cancer Evolution Consortium, Colin A Semple, Núria López-Bigas, Paul Flicek, Duncan T Odom, and Martin S Taylor. Pervasive lesion segregation shapes cancer genome evolution. *Nature*, 583(7815):265–270, July 2020.

[20] Matthew H Bailey, William U Meyerson, Lewis Jonathan Dursi, Liang-Bo Wang, Guanlan Dong, Wen-Wei Liang, Amila Weerasinghe, Shantao Li, Yize Li, Sean Kelso, MC3 Working Group, PCAWG novel somatic mutation calling methods working group, Gordon Saksena, Kyle Ellrott, Michael C Wendl, David A Wheeler, Gad Getz, Jared T Simpson, Mark B Gerstein, Li Ding, and PCAWG Consortium. Retrospective evaluation of whole exome and genome mutation calls in 746 cancer samples. *Nat. Commun.*, 11(1):4748, September 2020.

[21] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, March 2013.

[22] Maura Costello, Trevor J Pugh, Timothy J Fennell, Chip Stewart, Lee Lichtenstein, James C Meldrim, Jennifer L Fostel, Dennis C Friedrich, Danielle Perrin, Danielle Dionne, Sharon Kim, Stacey B Gabriel, Eric S Lander, Sheila Fisher, and Gad Getz. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative dna damage during sample preparation. *Nucleic Acids Res.*, 41(6):e67, April 2013.

[23] Lixin Chen, Pingfang Liu, Thomas C Evans, Jr, and Laurence M Ettwiller. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, 355(6326):752–756, February 2017.

[24] Federico Abascal, Luke M R Harvey, Emily Mitchell, Andrew R J Lawson, Stefanie V Lensing, Peter Ellis, Andrew J C Russell, Raul E Alcantara, Adrian Baez-Ortega, Yichen Wang, Eugene Jing Kwa, Henry Lee-Six, Alex Cagan, Tim H H Coorens, Michael Spencer Chapman, Sigurgeir Olafsson, Steven Leonard, David Jones, Heather E Machado, Megan Davies, Nina F Øbro, Krishnaa T Mahubani, Kieren Allinson, Moritz Gerstung, Kourosh Saeb-Parsy, David G Kent, Elisa Laurenti, Michael R Stratton, Raheleh Rahbari, Peter J Campbell, Robert J Osborne, and Iñigo Martincorena. Somatic mutation landscapes at single-molecule resolution. *Nature*, 593(7859):405–410, May 2021.

[25] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa,

H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[26] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, November 2008.

[27] 1000 Genomes Project Consortium, Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, November 2012.

[28] Justin Wagner, Nathan D Olson, Lindsay Harris, Jennifer McDaniel, Haoyu Cheng, Arkarachai Fungtammasan, Yih-Chii Hwang, Richa Gupta, Aaron M Wenger, William J Rowell, Ziad M Khan, Jesse Farek, Yiming Zhu, Aishwarya Pisupati, Medhat Mahmoud, Chunlin Xiao, Byunggil Yoo, Sayed Mohammad Ebrahim Sahraeian, Danny E Miller, David Jáspez, José M Lorenzo-Salazar, Adrián Muñoz-Barrera, Luis A Rubio-Rodríguez, Carlos Flores, Giuseppe Narzisi, Uday Shanker Evani, Wayne E Clarke, Joyce Lee, Christopher E Mason, Stephen E Lincoln, Karen H Miga, Mark T W Ebbert, Alaina Shumate, Heng Li, Chen-Shan Chin, Justin M Zook, and Fritz J Sedlazeck. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.*, 40(5):672–680, May 2022.

[29] K Osoegawa, A G Mammoser, C Wu, E Frengen, C Zeng, J J Catanese, and P J de Jong. A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.*, 11(3):483–496, March 2001.

[30] Erik Garrison, Jouni Sirén, Adam M Novak, Glenn Hickey, Jordan M Eizenga, Eric T Dawson, William Jones, Shilpa Garg, Charles Markello, Michael F Lin, Benedict Paten, and Richard Durbin. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.*, 36(9):875–879, October 2018.

[31] Valerie A Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A Kitts, Terence D Murphy, Kim D Pruitt, Françoise Thibaud-Nissen, Derek Albracht, Robert S Fulton, Milinn Kremitzki, Vincent Magrini, Chris Markovic, Sean McGrath, Karyn Meltz Steinberg, Kate Auger, William Chow, Joanna Collins, Glenn Harden, Timothy Hubbard, Sarah Pelan, Jared T Simpson, Glen Threadgold, James Torrance, Jonathan M Wood, Laura Clarke, Sergey Koren, Matthew Boitano, Paul Peluso, Heng Li, Chen-Shan Chin, Adam M Phillippy, Richard Durbin, Richard K Wilson, Paul Flicek, Evan E Eichler, and Deanna M Church. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, 27(5):849–864, May 2017.

[32] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Mikheenko Alla Bzikadze, Andrey V., Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow,

William Chow, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Erik Garrison, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Ying Sims, Daniela C. Soto, Ivan Sovic, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Aaron Wenger, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, April 2022.

[33] Sergey Aganezov, Stephanie M Yan, Daniela C Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J Taylor, Kishwar Shafin, Alaina Shumate, Chunlin Xiao, Justin Wagner, Jennifer McDaniel, Nathan D Olson, Michael E G Sauria, Mitchell R Vollger, Arang Rhie, Melissa Meredith, Skylar Martin, Joyce Lee, Sergey Koren, Jeffrey A Rosenfeld, Benedict Paten, Ryan Layer, Chen-Shan Chin, Fritz J Sedlazeck, Nancy F Hansen, Danny E Miller, Adam M Phillippy, Karen H Miga, Rajiv C McCoy, Megan Y Dennis, Justin M Zook, and Michael C Schatz. A complete reference genome improves analysis of human genetic variation. *Science*, 376(6588):eabl3533, April 2022.

[34] Michael A Lodato, Rachel E Rodin, Craig L Bohrson, Michael E Coulter, Alison R Barton, Minseok Kwon, Maxwell A Sherman, Carl M Vitzthum, Lovelace J Luquette, Chandri N Yandava, Pengwei Yang, Thomas W Chittenden, Nicole E Hatem, Steven C Ryu, Mollie B Woodworth, Peter J Park, and Christopher A Walsh. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, 359(6375):555–559, February 2018.

[35] Henry Lee-Six, Nina Friesgaard Øbro, Mairi S Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J Osborne, Brian J P Huntly, Inigo Martincorena, Elizabeth Anderson, Laura O'Neill, Michael R Stratton, Elisa Laurenti, Anthony R Green, David G Kent, and Peter J Campbell. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478, September 2018.

[36] Peter Ellis, Luiza Moore, Mathijs A Sanders, Timothy M Butler, Simon F Brunner, Henry Lee-Six, Robert Osborne, Ben Farr, Tim H H Coorens, Andrew R J Lawson, Alex Cagan, Mike R Stratton, Inigo Martincorena, and Peter J Campbell. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.*, 16(2):841–871, February 2021.

[37] P M Lizardi, X Huang, Z Zhu, P Bray-Ward, D C Thomas, and D C Ward. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.*, 19(3):225–232, July 1998.

[38] Fredrik Dahl, Johan Banér, Mats Gullberg, Maritha Mendel-Hartvig, Ulf Landegren, and Mats Nilsson. Circle-to-circle amplification for precise and sensitive DNA analysis. *Proc. Natl. Acad. Sci. U. S. A.*, 101(13):4548–4553, March 2004.

[39] Michael W Schmitt, Scott R Kennedy, Jesse J Salk, Edward J Fox, Joseph B Hiatt, and Lawrence A Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 109(36):14508–14513, September 2012.

[40] Margaret L Hoang, Isaac Kinde, Cristian Tomasetti, K Wyatt McMahon, Thomas A Rosenquist, Arthur P Grollman, Kenneth W Kinzler, Bert Vogelstein, and Nickolas Papadopoulos. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 113(35):9846–9851, August 2016.

[41] Iñigo Martincorena, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C Wedge, Anthony Fullam, Ludmil B Alexandrov, Jose M Tubio, Lucy Stebbings, Andrew Menzies, Sara Widaa, Michael R Stratton, Philip H Jones, and Peter J Campbell. Tumor evolution. high burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348(6237):880–886, May 2015.

[42] Young Seok Ju, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B Alexandrov, Raheleh Rahbari, David C Wedge, Helen R Davies, Manasa Ramakrishna, Anthony Fullam, Sancha Martin, Christopher Alder, Nikita Patel, Steve Gamble, Sarah O'Meara, Dilip D Giri, Torril Sauer, Sarah E Pinder, Colin A Purdie, Åke Borg, Henk Stunnenberg, Marc van de Vijver, Benita K T Tan, Carlos Caldas, Andrew Tutt, Naoto T Ueno, Laura J van 't Veer, John W M Martens, Christos Sotiriou, Stian Knappskog, Paul N Span, Sunil R Lakhani, Jórunn Erla Eyfjörd, Anne-Lise Børresen-Dale, Andrea Richardson, Alastair M Thompson, Alain Viari, Matthew E Hurles, Serena Nik-Zainal, Peter J Campbell, and Michael R Stratton. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543(7647):714–718, March 2017.

[43] Iñigo Martincorena, Joanna C Fowler, Agnieszka Wabik, Andrew R J Lawson, Federico Abascal, Michael W J Hall, Alex Cagan, Kasumi Murai, Krishnaa Mahbubani, Michael R Stratton, Rebecca C Fitzgerald, Penny A Handford, Peter J Campbell, Kourosh Saeb-Parsy, and Philip H Jones. Somatic mutant clones colonize the human esophagus with age. *Science*, 362(6417):911–917, November 2018.

[44] Simon F Brunner, Nicola D Roberts, Luke A Wylie, Luiza Moore, Sarah J Aitken, Susan E Davies, Mathijs A Sanders, Pete Ellis, Chris Alder, Yvette Hooks, Federico Abascal, Michael R Stratton, Inigo Martincorena, Matthew Hoare, and Peter J Campbell. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, 574(7779):538–542, October 2019.

[45] Henry Lee-Six, Sigurgeir Olafsson, Peter Ellis, Robert J Osborne, Mathijs A Sanders, Luiza Moore, Nikitas Georgakopoulos, Franco Torrente, Ayesha Noorani, Martin

Goddard, Philip Robinson, Tim H H Coorens, Laura O'Neill, Christopher Alder, Jingwei Wang, Rebecca C Fitzgerald, Matthias Zilbauer, Nicholas Coleman, Kourosh Saeb-Parsy, Inigo Martincorena, Peter J Campbell, and Michael R Stratton. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574(7779):532–537, October 2019.

[46] Kenichi Yoshida, Kate H C Gowers, Henry Lee-Six, Deepak P Chandrasekharan, Tim Coorens, Elizabeth F Maughan, Kathryn Beal, Andrew Menzies, Fraser R Millar, Elizabeth Anderson, Sarah E Clarke, Adam Pennycuick, Ricky M Thakrar, Colin R Butler, Nobuyuki Kakiuchi, Tomonori Hirano, Robert E Hynds, Michael R Stratton, Iñigo Martincorena, Sam M Janes, and Peter J Campbell. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, 578(7794):266–272, February 2020.

[47] Sigurgeir Olafsson, Rebecca E McIntyre, Tim Coorens, Timothy Butler, Hyunchul Jung, Philip S Robinson, Henry Lee-Six, Mathijs A Sanders, Kenneth Arestang, Claire Dawson, Monika Tripathi, Konstantina Strongili, Yvette Hooks, Michael R Stratton, Miles Parkes, Inigo Martincorena, Tim Raine, Peter J Campbell, and Carl A Anderson. Somatic evolution in non-neoplastic IBD-Affected colon. *Cell*, 182(3):672–684.e11, August 2020.

[48] Luiza Moore, Daniel Leongamornlert, Tim H H Coorens, Mathijs A Sanders, Peter Ellis, Stefan C Dentro, Kevin J Dawson, Tim Butler, Raheleh Rahbari, Thomas J Mitchell, Francesco Maura, Jyoti Nangalia, Patrick S Tarpey, Simon F Brunner, Henry Lee-Six, Yvette Hooks, Sarah Moody, Krishnaa T Mahbubani, Mercedes Jimenez-Linan, Jan J Brosens, Christine A Iacobuzio-Donahue, Inigo Martincorena, Kourosh Saeb-Parsy, Peter J Campbell, and Michael R Stratton. The mutational landscape of normal human endometrial epithelium. *Nature*, 580(7805):640–646, April 2020.

[49] Andrew R J Lawson, Federico Abascal, Tim H H Coorens, Yvette Hooks, Laura O'Neill, Calli Latimer, Keiran Raine, Mathijs A Sanders, Anne Y Warren, Krishnaa T A Mahbubani, Bethany Bareham, Timothy M Butler, Luke M R Harvey, Alex Cagan, Andrew Menzies, Luiza Moore, Alexandra J Colquhoun, William Turner, Benjamin Thomas, Vincent Gnanapragasam, Nicholas Williams, Doris M Rassl, Harald Vöhringer, Sonia Zumalave, Jyoti Nangalia, José M C Tubío, Moritz Gerstung, Kourosh Saeb-Parsy, Michael R Stratton, Peter J Campbell, Thomas J Mitchell, and Iñigo Martincorena. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, 370(6512):75–82, October 2020.

[50] Michael Spencer Chapman, Anna Maria Ranzoni, Brynelle Myers, Nicholas Williams, Tim H H Coorens, Emily Mitchell, Timothy Butler, Kevin J Dawson, Yvette Hooks, Luiza Moore, Jyoti Nangalia, Philip S Robinson, Kenichi Yoshida, Elizabeth Hook, Peter J Campbell, and Ana Cvejic. Lineage tracing of human development through somatic mutations. *Nature*, 595(7865):85–90, July 2021.

[51] Tim H H Coorens, Thomas R W Oliver, Rashesh Sanghvi, Ulla Sovio, Emma Cook, Roser Vento-Tormo, Muzlifah Haniffa, Matthew D Young, Raheleh Rahbari, Neil Sebire, Peter J Campbell, D Stephen Charnock-Jones, Gordon C S Smith, and Sam Behjati. Inherent mosaicism and extensive mutation of human placentas. *Nature*, 592(7852):80–85, April 2021.

[52] Philip S Robinson, Tim H H Coorens, Claire Palles, Emily Mitchell, Federico Abascal, Sigurgeir Olafsson, Bernard C H Lee, Andrew R J Lawson, Henry Lee-Six, Luiza Moore, Mathijs A Sanders, James Hewinson, Lynn Martin, Claudia M A Pinna, Sara Galavotti, Raheleh Rahbari, Peter J Campbell, Iñigo Martincorena, Ian Tomlinson, and Michael R Stratton. Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat. Genet.*, 53(10):1434–1442, October 2021.

[53] Sebastian Grossmann, Yvette Hooks, Laura Wilson, Luiza Moore, Laura O'Neill, Iñigo Martincorena, Thierry Voet, Michael R Stratton, Rakesh Heer, and Peter J Campbell. Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell*, 28(7):1262–1274.e5, July 2021.

[54] Luiza Moore, Alex Cagan, Tim H H Coorens, Matthew D C Neville, Rashesh Sanghvi, Mathijs A Sanders, Thomas R W Oliver, Daniel Leongamornlert, Peter Ellis, Ayesha Noorani, Thomas J Mitchell, Timothy M Butler, Yvette Hooks, Anne Y Warren, Mette Jorgensen, Kevin J Dawson, Andrew Menzies, Laura O'Neill, Calli Latimer, Mabel Teng, Ruben van Boxtel, Christine A Iacobuzio-Donahue, Inigo Martincorena, Rakesh Heer, Peter J Campbell, Rebecca C Fitzgerald, Michael R Stratton, and Raheleh Rahbari. The mutational landscape of human somatic and germline cells. *Nature*, 597(7876):381–386, September 2021.

[55] Seongyeol Park, Nanda Maya Mali, Ryul Kim, Jeong-Woo Choi, Junehawk Lee, Joonoh Lim, Jung Min Park, Jung Woo Park, Donghyun Kim, Taewoo Kim, Kijong Yi, June Hyug Choi, Seong Gyu Kwon, Joo Hee Hong, Jeonghwan Youk, Yohan An, Su Yeon Kim, Soo A Oh, Youngoh Kwon, Dongwan Hong, Moonkyu Kim, Dong Sun Kim, Ji Young Park, Ji Won Oh, and Young Seok Ju. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature*, 597(7876):393–397, September 2021.

[56] Stanley W K Ng, Foad J Rouhani, Simon F Brunner, Natalia Brzozowska, Sarah J Aitken, Ming Yang, Federico Abascal, Luiza Moore, Efterpi Nikitopoulou, Lia Chappell, Daniel Leongamornlert, Aleksandra Ivovic, Philip Robinson, Timothy Butler, Mathijs A Sanders, Nicholas Williams, Tim H H Coorens, Jon Teague, Keiran Raine, Adam P Butler, Yvette Hooks, Beverley Wilson, Natalie Birtchnell, Huw Naylor, Susan E Davies, Michael R Stratton, Iñigo Martincorena, Raheleh Rahbari, Christian Frezza, Matthew Hoare, and Peter J Campbell. Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature*, 598(7881):473–478, October 2021.

[57] Aaron M Newman, Alexander F Lovejoy, Daniel M Klass, David M Kurtz, Jacob J Chabon, Florian Scherer, Henning Stehr, Chih Long Liu, Scott V Bratman, Carmen Say, Li Zhou, Justin N Carter, Robert B West, George W Sledge, Joseph B Shrager, Billy W Loo, Jr, Joel W Neal, Heather A Wakelee, Maximilian Diehn, and Ash A Alizadeh. Integrated digital error suppression for improved detection of circulating tumor DNA. *Nat. Biotechnol.*, 34(5):547–555, May 2016.

[58] M J Levene, J Korlach, S W Turner, M Foquet, H G Craighead, and W W Webb. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607):682–686, January 2003.

[59] Jonas Korlach, Patrick J Marks, Ronald L Cicero, Jeremy J Gray, Devon L Murphy, Daniel B Roitman, Thang T Pham, Geoff A Otto, Mathieu Foquet, and Stephen W Turner. Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U. S. A.*, 105(4):1176–1181, January 2008.

[60] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, January 2009.

[61] Jonas Korlach, Arek Bibillo, Jeffrey Wegener, Paul Peluso, Thang T Pham, Insil Park, Sonya Clark, Geoff A Otto, and Stephen W Turner. Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides Nucleotides Nucleic Acids*, 27(9):1072–1083, September 2008.

[62] Benjamin A Flusberg, Dale R Webster, Jessica H Lee, Kevin J Travers, Eric C Olivares, Tyson A Clark, Jonas Korlach, and Stephen W Turner. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods*, 7(6):461–465, June 2010.

[63] Tyson A Clark, Kristi E Spittle, Stephen W Turner, and Jonas Korlach. Direct detection and sequencing of damaged DNA bases. *Genome Integr.*, 2:10, December 2011.

[64] Aaron M Wenger, Paul Peluso, William J Rowell, Pi-Chuan Chang, Richard J Hall, Gregory T Concepcion, Jana Ebler, Arkarachai Fungtammasan, Alexey Kolesnikov, Nathan D Olson, Armin Töpfer, Michael Alonge, Medhat Mahmoud, Yufeng Qian, Chen-Shan Chin, Adam M Phillippy, Michael C Schatz, Gene Myers, Mark A DePristo, Jue Ruan, Tobias Marschall, Fritz J Sedlazeck, Justin M Zook, Heng Li, Sergey Koren, Andrew Carroll, David R Rank, and Michael W Hunkapiller. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, 37(10):1155–1162, October 2019.

[65] Jeffrey A Bailey, Zhiping Gu, Royden A Clark, Knut Reinert, Rhea V Samonte, Stuart Schwartz, Mark D Adams, Eugene W Myers, Peter W Li, and Evan E Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, August 2002.

[66] H F Willard. Chromosome-specific organization of human alpha satellite DNA. *Am. J. Hum. Genet.*, 37(3):524–532, May 1985.

[67] Helen Skaletsky, Tomoko Kuroda-Kawaguchi, Patrick J Minx, Holland S Cordum, Ladeana Hillier, Laura G Brown, Sjoerd Repping, Tatyana Pyntikova, Johar Ali, Tamberlyn Bieri, Asif Chinwalla, Andrew Delehaunty, Kim Delehaunty, Hui Du, Ginger

Fewell, Lucinda Fulton, Robert Fulton, Tina Graves, Shun-Fang Hou, Philip Latrielle, Shawn Leonard, Elaine Mardis, Rachel Maupin, John McPherson, Tracie Miner, William Nash, Christine Nguyen, Philip Ozersky, Kymberlie Pepin, Susan Rock, Tracy Rohlfing, Kelsi Scott, Brian Schultz, Cindy Strong, Aye Tin-Wollam, Shiaw-Pyng Yang, Robert H Waterston, Richard K Wilson, Steve Rozen, and David C Page. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942):825–837, June 2003.

[68] Yu Lin, Jeffrey Yuan, Mikhail Kolmogorov, Max W Shen, Mark Chaisson, and Pavel A Pevzner. Assembly of long error-prone reads using de bruijn graphs. *Proc. Natl. Acad. Sci. U. S. A.*, 113(52):E8396–E8405, December 2016.

[69] Eugene W Myers. The fragment assembly string graph. *Bioinformatics*, 21 Suppl 2:ii79–85, September 2005.

[70] Chen-Shan Chin, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, Ronan O'Malley, Rosa Figueroa-Balderas, Abraham Morales-Cruz, Grant R Cramer, Massimo Delledonne, Chongyuan Luo, Joseph R Ecker, Dario Cantu, David R Rank, and Michael C Schatz. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*, 13(12):1050–1054, December 2016.

[71] Sergey Koren, Brian P Walenz, Konstantin Berlin, Jason R Miller, Nicholas H Bergman, and Adam M Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, 27(5):722–736, May 2017.

[72] Ali Bashir, Aaron Klammer, William P Robins, Chen-Shan Chin, Dale Webster, Ellen Paxinos, David Hsu, Meredith Ashby, Susana Wang, Paul Peluso, Robert Sebra, Jon Sorenson, James Bullard, Jackie Yen, Marie Valdovino, Emilia Mollova, Khai Luong, Steven Lin, Brianna LaMay, Amruta Joshi, Lori Rowe, Michael Frace, Cheryl L Tarr, Maryann Turnsek, Brigid M Davis, Andrew Kasarskis, John J Mekalanos, Matthew K Waldor, and Eric E Schadt. A hybrid approach for the automated finishing of bacterial genomes. *Nat. Biotechnol.*, 30(7):701–707, July 2012.

[73] Chen-Shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, 10(6):563–569, June 2013.

[74] John Huddleston, Swati Ranade, Maika Malig, Francesca Antonacci, Mark Chaisson, Lawrence Hon, Peter H Sudmant, Tina A Graves, Can Alkan, Megan Y Dennis, Richard K Wilson, Stephen W Turner, Jonas Korlach, and Evan E Eichler. Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.*, 24(4):688–696, April 2014.

[75] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, Heng Dai, Markus Hsi-Yang Fritz, Han Cao, Ariella Cohain, Gintaras

Deikus, Russell E Durrett, Scott C Blanchard, Roger Altman, Chen-Shan Chin, Yan Guo, Ellen E Paxinos, Jan O Korbel, Robert B Darnell, W Richard McCombie, Pui-Yan Kwok, Christopher E Mason, Eric E Schadt, and Ali Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, 12(8):780–786, August 2015.

[76] Olga Dudchenko, Sanjit S Batra, Arina D Omer, Sarah K Nyquist, Marie Hoeger, Neva C Durand, Muhammad S Shamim, Ido Machol, Eric S Lander, Aviva Presser Aiden, and Erez Lieberman Aiden. De novo assembly of the aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, 356(6333):92–95, April 2017.

[77] James T Robinson, Douglass Turner, Neva C Durand, Helga Thorvaldsdóttir, Jill P Mesirov, and Erez Lieberman Aiden. Juicebox.js provides a Cloud-Based visualization system for Hi-C data. *Cell Syst*, 6(2):256–258.e1, February 2018.

[78] Olga Dudchenko, Muhammad S Shamim, Sanjit S Batra, Neva C Durand, Nathaniel T Musial, Ragib Mostofa, Melanie Pham, Brian Glenn St Hilaire, Weijie Yao, Elena Stamenova, Marie Hoeger, Sarah K Nyquist, Valeriya Korchina, Kelcie Pletch, Joseph P Flanagan, Ania Tomaszewicz, Denise McAloose, Cynthia Pérez Estrada, Ben J Novak, Arina D Omer, and Erez Lieberman Aiden. The juicebox assembly tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. *bioRxiv*, January 2018.

[79] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy P L Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.*, October 2018.

[80] David Porubsky, Peter Ebert, Peter A Audano, Mitchell R Vollger, William T Harvey, Pierre Marijon, Jana Ebler, Katherine M Munson, Melanie Sorensen, Arvis Sulovari, Marina Haukness, Maryam Ghareghani, Human Genome Structural Variation Consortium, Peter M Lansdorp, Benedict Paten, Scott E Devine, Ashley D Sanders, Charles Lee, Mark J P Chaisson, Jan O Korbel, Evan E Eichler, and Tobias Marschall. Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol.*, 39(3):302–308, March 2021.

[81] Benjamin J Matthews, Olga Dudchenko, Sarah B Kingan, Sergey Koren, Igor Antoshechkin, Jacob E Crawford, William J Glassford, Margaret Herre, Seth N Redmond, Noah H Rose, Gareth D Weedall, Yang Wu, Sanjit S Batra, Carlos A Brito-Sierra, Steven D Buckingham, Corey L Campbell, Saki Chan, Eric Cox, Benjamin R Evans, Thanyalak Fansiri, Igor Filipović, Albin Fontaine, Andrea Gloria-Soria, Richard Hall, Vinita S Joardar, Andrew K Jones, Raissa G G Kay, Vamsi K Kodali, Joyce Lee, Gareth J Lycett, Sara N Mitchell, Jill Muehling, Michael R Murphy, Arina D Omer, Frederick A Partridge, Paul Peluso, Aviva Presser Aiden, Vidya Ramasamy, Gordana Rašić, Sourav Roy, Karla Saavedra-Rodriguez, Shruti Sharan, Atashi Sharma, Melissa Laird Smith, Joe Turner, Allison M Weakley, Zhilei Zhao, Omar S Akbari, William C Black, 4th, Han Cao, Alistair C Darby, Catherine A Hill, J Spencer Johnston, Terence D Murphy, Alexander S Raikhel, David B Sattelle, Igor V Sharakhov, Bradley J White, Li Zhao, Erez Lieberman Aiden, Richard S Mann, Louis Lambrechts, Jeffrey R Powell,

Maria V Sharakhova, Zhijian Tu, Hugh M Robertson, Carolyn S McBride, Alex R Hastie, Jonas Korlach, Daniel E Neafsey, Adam M Phillippy, and Leslie B Vosshall. Improved reference genome of aedes aegypti informs arbovirus vector control. *Nature*, 563(7732):501–507, November 2018.

[82] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, Sunir Malla, Hannah Marriott, Tom Nieto, Justin O'Grady, Hugh E Olsen, Brent S Pedersen, Arang Rhie, Hollian Richardson, Aaron R Quinlan, Terrance P Snutch, Louise Tee, Benedict Paten, Adam M Phillippy, Jared T Simpson, Nicholas J Loman, and Matthew Loose. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, 36(4):338–345, April 2018.

[83] Miten Jain, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson, and Karen H Miga. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.*, 36(4):321–323, April 2018.

[84] Karen H Miga, Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F Willard, and W James Kent. Centromere reference models for human chromosomes x and y satellite arrays. *Genome Res.*, 24(4):697–707, April 2014.

[85] Chen-Shan Chin. Human genome assembly in 100 minutes. *bioRxiv*, 2019.

[86] Sergey Nurk, Brian P Walenz, Arang Rhie, Mitchell R Vollger, Glennis A Logsdon, Robert Grothe, Karen H Miga, Evan E Eichler, Adam M Phillippy, and Sergey Koren. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.*, 30(9):1291–1305, September 2020.

[87] Haoyu Cheng, Gregory T Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, 18(2):170–175, February 2021.

[88] Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, 12(5):363–376, May 2011.

[89] Mark J P Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, Jane M Landolin, John A Stamatoyannopoulos, Michael W Hunkapiller, Jonas Korlach, and Evan E Eichler. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611, January 2015.

[90] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, 15(6):461–468, June 2018.

[91] Luca Denti, Parsoa Khorsand, Paola Bonizzoni, Fereydoun Hormozdiari, and Rayan Chikhi. SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nat. Methods*, December 2022.

[92] Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.*, 14(2):125–138, February 2013.

[93] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3D genome. *Nat. Rev. Genet.*, 19(7):453–467, July 2018.

[94] Jan O Korbel and Peter J Campbell. Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152(6):1226–1236, March 2013.

[95] Sylvan C Baca, Davide Prandi, Michael S Lawrence, Juan Miguel Mosquera, Alessandro Romanel, Yotam Drier, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Mahmoud Ghandi, Eliezer Van Allen, Gregory V Kryukov, Andrea Sboner, Jean-Philippe Theurillat, T David Soong, Elizabeth Nickerson, Daniel Auclair, Ashutosh Tewari, Himisha Beltran, Robert C Onofrio, Gunther Boysen, Candace Guiducci, Christopher E Barbieri, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Gordon Saksena, Douglas Voet, Alex H Ramos, Wendy Winckler, Michelle Cipicchio, Kristin Ardlie, Philip W Kantoff, Michael F Berger, Stacey B Gabriel, Todd R Golub, Matthew Meyerson, Eric S Lander, Olivier Elemento, Gad Getz, Francesca Demichelis, Mark A Rubin, and Levi A Garraway. Punctuated evolution of prostate cancer genomes. *Cell*, 153(3):666–677, April 2013.

[96] Amy Marie Yu and Mitch McVey. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res.*, 38(17):5706–5717, September 2010.

[97] Zhi-Dong Zhou, Joseph Jankovic, Tetsuo Ashizawa, and Eng-King Tan. Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. *Nat. Rev. Neurol.*, 18(3):145–157, March 2022.

[98] Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res.*, 38(15):e159, August 2010.

[99] Nucleotide sequence of bacteriophage phix174 dna.

[100] Mark J Chaisson and Glenn Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *BMC Bioinformatics*, 13(238), September 2012.

[101] Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, Kevin J Dawson, Nicole Mende, Emily F Calderbank, Hyunchul Jung, Thomas Mitchell, Tim H H Coorens, David H Spencer, Heather Machado, Henry Lee-Six, Megan Davies, Daniel Hayler, Margarete A Fabre, Krishnaa Mahbubani, Federico Abascal, Alex Cagan, George S Vassiliou, Joanna Baxter, Inigo Martincorena, Michael R Stratton, David G Kent, Krishna Chatterjee, Kourosh Saeb Parsy, Anthony R Green, Jyoti Nangalia, Elisa Laurenti, and Peter J Campbell. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606(7913):343–350, June 2022.

[102] Mia Petljak, Ludmil B Alexandrov, Jonathan S Brammeld, Stacey Price, David C Wedge, Sebastian Grossmann, Kevin J Dawson, Young Seok Ju, Francesco Iorio, Jose M C Tubio, Ching Chiek Koh, Ilias Georgakopoulos-Soares, Bernardo Rodríguez-Martín, Burçak Otlu, Sarah O'Meara, Adam P Butler, Andrew Menzies, Shriram G Bhosle, Keiran Raine, David R Jones, Jon W Teague, Kathryn Beal, Calli Latimer, Laura O'Neill, Jorge Zamora, Elizabeth Anderson, Nikita Patel, Mark Maddison, Bee Ling Ng, Jennifer Graham, Mathew J Garnett, Ultan McDermott, Serena Nik-Zainal, Peter J Campbell, and Michael R Stratton. Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis. *Cell*, 176(6):1282–1294.e20, March 2019.

[103] Sheina B Sim, Renee L Corpuz, Tyler J Simmonds, and Scott M Geib. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics*, 23(1):157, February 2022.

[104] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, September 2018.

[105] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[106] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, 36(10):983–987, November 2018.

[107] Heng Li. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, 27(5):718–719, March 2011.

[108] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, November 2011.

[109] Fernando G Osorio, Axel Rosendahl Huber, Rurika Oka, Mark Verheul, Sachin H Patel, Karlijn Hasaart, Lisanne de la Fonteijne, Ignacio Varela, Fernando D Camargo, and Ruben van Boxtel. Somatic mutations reveal lineage relationships and Age-Related mutagenesis in human hematopoiesis. *Cell Rep.*, 25(9):2308–2316.e4, November 2018.

[110] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498, May 2011.

[111] Sangtae Kim, Konrad Scheffler, Aaron L Halpern, Mitchell A Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Doruk Beyter, Peter Krusche, and Christopher T Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods*, 15(8):591–594, August 2018.

[112] pysam developers. pysam.

[113] Lianming Du, Qin Liu, Zhenxin Fan, Jie Tang, Xiuyue Zhang, Megan Price, Bisong Yue, and Kelei Zhao. Pyfastx: a robust python package for fast random access to sequences from plain and gzipped FASTA/Q files. *Brief. Bioinform.*, 22(4), July 2021.

[114] Brent S Pedersen and Aaron R Quinlan. cyvcf2: fast, flexible variant analysis with python. *Bioinformatics*, 33(12):1867–1869, June 2017.

[115] Yan Gao, Yongzhuang Liu, Yanmei Ma, Bo Liu, Yadong Wang, and Yi Xing. abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *Bioinformatics*, 37(15):2209–2211, August 2021.