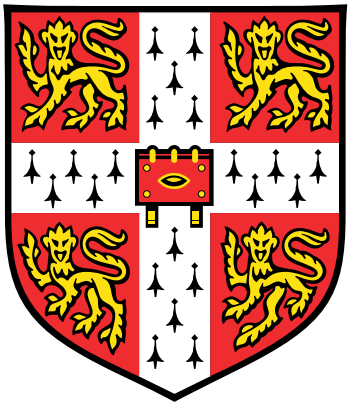
Germline and somatic mutational processes across the Tree of Life



Sangjin Lee

Wellcome Trust Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Downing College

April 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
April 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

Pacific Biosciences' circular consensus sequencing uses an engineered DNA polymerase with high processivity to repeatedly sequence the forward and reverse strand of a circular template. Leveraging redundancies and complementary base pairing between the forward and reverse strand, a highly accurate circular consensus sequence (CCS) read with at least Q20 read accuracy is generated.

In this PhD thesis, I propose a hypothesis based on the similarities between duplex and CCS library preparation protocols. I conjecture that a CCS read is as accurate or more accurate than a duplex read, which is often used for ultra-rare somatic mutation detection. Using cord blood granulocytes with few somatic mutations, I demonstrate that a subset of CCS bases has higher base accuracy than duplex bases. In addition, I empirically calculate the CCS base accuracy to range from Q60 to Q90 depending on the substitution error and the trinucleotide sequence context, which is sufficiently accurate to enable single molecule somatic mutation detection in human samples and potentially in non-human samples with an unknown somatic mutation rate.

To enable the somatic mutation detection and analysis across the tree of life, I design and develop a tool called himut to detect and phase somatic mutations from bulk normal tissue using CCS reads. I select a set of samples with a single somatic mutational process to distinguish recently acquired somatic mutations from false positive substitutions arising from sequencing and systematic bioinformatics errors. In addition, I ascertain that inaccurate estimation of base quality scores during CCS generation as one of the primary causes of false positive substitution detection.

The Darwin Tree of Life (DToL) project aspires to sequence and generate reference genomes for around 70,000 eukaryotic species in Great Britain and Ireland. To date, the DToL consortium has publicly released CCS reads and reference genomes for approximately 600 eukaryotic species. I use himut to detect somatic mutations across the tree of life from DToL datasets and perform *de novo* mutational signature extraction to identify new and conserved somatic mutational processes. One striking observation is the episodic emergence and establishment of somatic mutational processes. The conservation of C>T

somatic mutations at CG dinucleotides, resulting from spontaneous deamination of 5-methylcytosine to thymine, in animal, fungi and plant kingdom is one such example. I believe that methods described in this PhD thesis will facilitate the discovery and analysis of somatic mutational processes and enhance our understanding of natural selection.

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

1.1 Somatic mutations

1.1.1 Somatic mutation detection

1.1.2 Somatic mutations in normal tissues

1.1.3 Somatic mutational processes

1.1.4 Mutational signature analysis

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, epigenetic, 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].

1.1.5 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.

1.1.6 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 different 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

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.

1.3.1 Origins of Life, Prebiotic Earth, RNA world, 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 [].

1.4 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

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]

Considering the similarities between duplex and CCS sequencing and based on my understanding of the duplex sequencing method [39, 40] and the recently developed nanorate sequencing protocol [24], 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.

2.1.1 Sanger sequencing

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, what is in essence, a duplex read 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* [].

2.1.2 Duplex sequencing

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. In addition, repeated Sanger sequencing of the same template molecule could have produced duplex reads with similar performance, but the sequencing cost would have been impractical.

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 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 taking advantage of 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 in identifying library and amplification 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].

2.1.3 Nanorate sequencing

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 terminate 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.

2.1.4 Circular consensus sequencing

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. I would also like to emphasise that the commercial CCS library preparation protocol is more similar to the original duplex sequencing protocol than the nanorate library protocol. 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 higher base accuracy allows more recently diverged repeats to be distinguished from each other in combination [].

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 me 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. My understanding of somatic mutational processes across different tissue types was critical in selecting the samples for single molecule somatic mutation detection evaluation and demonstration 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 assess the potential for single molecule somatic mutation detection using CCS reads. 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 normal 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 functions as part of the innate immune response to viruses and retrotransposons. APOBEC enzymes act upon single-stranded DNA and RNA as cytidine deaminase and catalyse 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 is 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 [].

During eukaryotic DNA replication, DNA polymerase δ (POLD) and ε (POLE) are critical for DNA synthesis on the lagging and leading strand, respectively []. POLD and POLE enzymes both have intrinsic proofreading capabilities and their 3'-5' exonuclease activity removes the 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 introduce mutations at a constant rate throughout life and hence, the 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. Somatic mutation rate in normal granulocytes has been determined to be approximately 16.8 substitutions per cell per year and 0.71 indel per cell per year []. 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, I first assessed the accuracy of the BQ score estimate using CCS reads from cord blood granulocytes. The number of matches and mismatches were counted for each BQ score estimate to calculate the empirical BQ score. I considered reference allele and germline SNPs as matches and all other SBS as mismatches. Germline mutation detection using himut is described below. I 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. I, thereafter, calculated empirical BQ for each BQ score estimate (eq. 4.1):

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

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 [109, 110]. 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 (eq. 4.2):

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

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 (eq. 4.3):

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

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 (eq. 4.4):

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

where $P(b|G)$ is the probability of observing the base given the genotype and is defined as such (eq. 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 alleles 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 (eq. 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 a 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 misalignment of residual adapter sequences might generate somatic mutation candidates. Therefore, candidates located near start and ends of reads are filtered as specified with the minimum trimming parameter.

8. The number of mismatches adjacent to the candidate needs to be below the maximum mismatch count threshold within a given mismatch window as an alignment error can be mistaken as a somatic mutation.

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 [111], pyfastx [112] and cyvcf2 [113], respectively. In addition, multiprocessing Python package [] was used to enable parallel processing of each chromosomes.

2.2.5 Panel of Normal construction

I 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. I 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, the power to distinguish errors from somatic mutations will also increase in the future.

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. I 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 with germline mutations 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 the minimum haplotype count threshold to select regions where both haplotypes have been sampled sufficiently and to prevent misclassification of hetSNPs as somatic mutations.

2.2.8 Somatic mutation count normalisation

To normalise the number of substitutions per trinucleotide sequence context, the SBS96 classification system and the same conditions as somatic mutation detection is used to the calculate of the number of callable reference and CCS bases. I would like to highlight that only the reference bases where homozygous reference allele has been called without an indel will considered as a callable reference base.

Under the SBS96 classification, SBS is categorised according to 6 possible substitution types in the pyrimidine context (C>A, C>G, C>T, T>A, T>C and T>G) and 16 possible trinucleotide sequence context derived from the 4 possible bases upstream and downstream of the substitution.

The frequency of each trinucleotide is calculated for the reference f_i^g, callable reference $f_i^{g_{\text{callable}}}$, and callable CCS f_i^{CCS} bases from the reference genome FASTA file, the number of callable reference bases and the number of callable CCS bases, respectively (eq. 2.8).

$$f_i = \frac{t_i}{\sum_{i=1}^{32} t_i} \tag{2.8}$$

where t denote a specific trinucleotide. There are 32 possible trinucleotide sequence contexts where pyrimidine is the middle base.

The number of somatic mutations is, thereafter, normalised with the ratio of reference to callable reference trinucleotide frequency and the ratio of callable reference and callable CCS base trinucleotide frequency according to the substitution and the trinucleotide sequence context. ACA>A somatic mutation count, for example, is normalised as follows (eq. 2.9):

$$S'_{\text{ACA>A}} = S_{\text{ACA>A}} \times r^{g}_{\text{ACA}} \times r^{\text{callable}}_{\text{ACA}} \tag{2.9}$$

where $S'_{\text{ACA>A}}$ is the normalised substitution count, $S_{\text{ACA>A}}$ is the raw substitution count, r^{g}_{ACA} is the ratio of reference to callable reference ACA frequency and $r^{\text{callable}}_{\text{ACA}}$ is the ratio of callable reference and CCS base ACA frequency.

The ratio of reference to callable reference trinucleotide frequency (eq. 2.10) and the ratio of callable reference and CCS base trinucleotide frequency is calculated as follows (eq. 2.11):

$$r^{g}_{i} = \frac{f_i^{g_{\text{callable}}}}{f_i^{g}} \tag{2.10}$$

$$r^{\text{callable}}_{i} = \frac{f_i^{g_{\text{callable}}}}{f_i^{\text{CCS}_{\text{callable}}}} \tag{2.11}$$

2.2.9 Mutation burden calculation

The somatic mutation rate is calculated for each trinucleotide sequence context from the normalised somatic mutation counts and the number of callable CCS bases (eq. 2.12)

$$m_{\text{ACA}} = \frac{S'_{\text{ACA>C}} + S'_{\text{ACA>G}} + S'_{\text{ACA>T}}}{t^{\text{CCS}_{\text{callable}}}_{\text{ACA}}} \tag{2.12}$$

where m_{ACA} is the somatic mutation rate for the ACA trinucleotide sequence context.

To calculate the mutation burden per cell, genomic mutation burden is first calculated using the trinucleotide sequence context specific somatic mutation rate and the number

of trinucleotides in the reference FASTA file and the genomic mutation burden is adjusted with the ploidy of the sample to derive the mutation burden per cell (eq. 2.13)

$$g_{\text{burden}} = n * (\sum_{i=1}^{32} m_i * t_i^g) \tag{2.13}$$

where n is the ploidy of the sample, m_i is the trinucleotide sequence context somatic mutation rate and t_i^g is the number of trinucleotides in the reference genome.

2.2.10 CCS error rate per trinucleotide sequence context

To calculate the substitution error rate per trinucleotide sequence context, cord blood CCS reads were regenerated with 10 full-length subreads per CCS read using pbccs. The first subread and the last subread were excluded from CCS read generation. The number of subreads were set as a constant as the number of subreads influenced the CCS BQ score estimate and as the increase in number of subreads per CCS reads decreased the accuracy of the BQ score estimate (discussed and demonstrated later in Chapter 3). Cord blood CCS reads were subsequently processed as described above for read alignment, somatic mutation detection and somatic mutation count normalisation.

Somatic mutations detected from the cord blood granulocyte are attributable to the clock-like mutational process and non-biological processes (library, sequencing, and alignment errors). As the number of called somatic mutations greatly exceeds the expected number of somatic mutations of 40-50 somatic mutations per cell [114, 101], the cord blood sample can be assumed to be enriched with false positive mutations.

To accurately ascertain the number of false positive mutations, the number of true positive somatic mutations were first estimated from the number of callable bases and the cord blood somatic mutational process [101] and were subtracted from the normalised somatic mutation counts. The number of normalised false positive somatic mutation counts, and the number of callable bases were subsequently used to estimate the substitution error rate per trinucleotide sequence context.

2.2.11 CCS base quality score recalibration

ACTC [] was used to align subreads to CCS reads from the same ZMW to determine the sequence orientation of subreads and to exclude subreads produced from erroneous adapter sequence detection. DeepConsensus accepts as input the BAM file where subreads have been aligned to CCS read from the same ZMW, 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 at least 10 subreads. CCS bases with unanimous support were assigned Q93 BQ score while all other bases were assigned Q0 BQ score. Himut was, thereafter, used to call somatic mutations from abPOA andDeepConsensus polished CCS reads to assess the impact of BQ score recalibration.

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.

I 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 my method.

I, 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

	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

Table 2.1 Experimental Data

DNA molecules (e.g, failed ligation products) not suitable for sequencing. 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 poly(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 new 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 resulting from internal priming and early 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 error-free 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 consensus sequence is constructed from multiple sequence alignment of subreads and is polished based on the realignment of subreads back to the draft consensus 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 thought to be randomly introduced without sequence context bias and are independent of each other. 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 between 50% of median subread length and 200% of median subread length 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 shorter subreads and multiple subreads can be concatenated into a single long 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). I observed that X% of ZMWs have fragmented and/or concatenated subreads (Figure **??**). I 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, I 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, hence, 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. Clonal somatic mutations, hence, can often be confused with heterozygous germline mutations in normal tissues. Paired tumour-normal sequencing, therefore, is often performed to distinguish germline mutations from somatic mutations in a tumour sample. Here, I present how himut distinguishes errors and germline mutations from somatic mutations in bulk normal tissue, leveraging CCS read length and base accuracy.

I, first, compared germline SNPs detected from both himut and deepvariant to assess whether our algorithm can accurately call germline mutations (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. I believe that algorithmic differences account for disparities in the number of SNPs called with himut and deep-

variant, 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.

Chapter 3

Germline and somatic mutational processes across the tree of life

Both in space and time, we seem to be brought somewhat near to that great fact—the mystery of mysteries—the first appearance of new beings on this earth []

[Charles Darwin]

3.1 Introduction

Somatic mutations can occur in cells at all stages of life and in all tissues.

3.1.1 The Darwin Tree of Life Project

3.1.2 CCS sequencing and *de novo* assembly

3.2 Results

3.2.1 Somatic mutation detection

3.2.1.1 Phorcus lineatus somatic mutation rate

3.2.2 Germline and somatic mutational processes

3.3 Conclusion

3.4 Materials and Methods

3.4.1 CCS library preparation, sequencing and *de novo* assembly

To date, the DToL consortium has collected, prepared, and sequenced approximately ~3000 eukaryotic samples in Great Britain and Ireland. In addition, reference genomes for around 600 eukaryotic species have been assembled and made available to the public, which is accompanied by a genome note that details the process from sample acquisition to chromosome-length scaffold construction.

The DToL project initially used a combination of sequencing (CLR, CCS and linked reads) and scaffolding (e.g. Hi-C reads and BioNano genome maps) technologies to generate chromosome-length scaffolds. The DToL project currently uses HiFiAdapterFilter [] to remove CCS reads with adapter sequences, either hifiasm [] or hicanu [] for de novo assembly of contigs from CCS reads, purgedups [] to remove haplotype duplication, arrow [] for contig polishing and SALSA [] to order and orient contigs into chromosome-length scaffolds with Hi-C reads. If both the parent and child was sequenced, trio-canu was used to generate haplotype phased contigs. The chromosome-length scaffolds are, thereafter, manually curated using a Hi-C contact matrix to identify and correct misassemblies and to perform additional scaffolding where appropriate. If transcriptome data was available through either RNA or isoform sequencing, gene annotation was also performed in collaboration with the EMBL-EBI eukaryotic annotation team. The specific method and algorithm described here is subject to change with updates to the sequencing method, *de novo* assembly and scaffolding algorithm.

3.4.2 *Phorcus lineatus* somatic mutation rate measurement

To calculate the somatic mutation rate of *P. lineatus* (thick top shell), samples of different ages (3, 5 and 15) were collected from Plymouth, UK. Collaborators at the Marine Biological Association (MBA) determined the age of the samples from growth marks on the shells of samples. As recommended, a bench-mounted vice was first used to crush the shell and to carefully separate the sample from the shell (personal communication with Robert Mrowicki at the MBA). In addition, disposable scalpels were used during the dissection to prevent cross-contamination between the samples. HMW DNA was subsequently extracted from the foot muscle using the Circulomics Nanobind Tissue Big DNA Kit (SKU 102-302-100). CCS libraries were prepared following the low-input CCS library preparation protocol () and BluePippin system () was used to size select CCS libraries prior to sequencing.

CCS BQ score is a function of the number of supporting subreads and the concordance between the CCS base and subread bases. The DNAP processivity and CCS read length determine the number of full-length subreads per CCS read, which in turn influences the number of Q93 CCS bases from which potential somatic mutations can be identified. To account for the differences in the number of subreads per CCS read for each ZMW, the raw subreads BAM file was parsed using a custom script to select 10 full-length subreads per ZMW. The script calculates the median subread length for each ZMW and considers subreads between 0.9 times the median subread length and 1.1 times the median subread length as full-length subreads. The processed subreads BAM file was, thereafter, provided as an input to the pbccs algorithm to re-generate CCS reads. CCS reads were subsequently processed as described in chapter 2 and below. Except for the *P. lineatus* somatic mutation rate measurement, CCS reads generated with default pbccs parameters for used for the rest of the analysis.

3.4.3 Germline and somatic mutation detection

As detailed in chapter 2, CCS reads with adapter sequences were identified using HiFi-AdapterFilt [] and subsequently discarded. In addition, if ultra-low input CCS library preparation protocol was used for CCS generation and if this was documented, CCS reads were also excluded from downstream sequence analysis (this information, however, was not always available). CCS reads were, thereafter, aligned to the assembled reference genomes using minimap2 [] and primary alignments were selected, sorted and merged into a single BAM file using samtools []. Germline mutations were called using deepvariant

[]. Somatic mutations were detected, and mutation burden was calculated using himut with default parameters. In addition, heterozygous SNPs were phased to construct haplotype blocks using himut, and haplotype phased CCS reads were subsequently used to call haplotype phased somatic mutations where applicable. Somatic mutation detection was again restricted to the autosomes of the reference genome.

As CCS reads and reference genomes are derived from the same sample, homozygous germline mutations indicate assembly errors and analysis of germline mutations are restricted to heterozygous mutations for samples with a diploid genome.

The detection of somatic mutations across the tree of life followed a similar approach to the one described in chapter 2, but with minor modifications. When somatic mutations were called from DToL eukaryotic species, a VCF file containing germline mutations was supplied to himut to calculate heterozygosity (θ) and the genotype prior $P(G)$. In addition, because a single sample was sequenced per species and as population-scale sequencing studies has not been performed for these species, PoN VCF file could not be generated and VCF file with common SNPs were not available for distinguishing false positive substitutions arising from systematic errors and gDNA contamination, respectively. However, given that CCS library and reference genome originate from the same sample, false positive substitutions arising from alignment errors should be minimal and CCS reads resulting from gDNA contamination should be excluded from the analysis based on their sequence identity.

3.4.4 Mutational signature extraction and analysis

As described in chapter 1, there are 6 substitution types (C>A, C>G, C>T, T>A, T>C, T>G) in the pyrimidine context and 16 trinucleotide sequence contexts for each substitution class, creating the canonical SBS96 classification system. Since the ancestral allele is known for somatic mutations, the SBS96 classification system is often used to categorise somatic substitutions. In contrast, because the ancestral allele is unknown for germline mutations, the SBS52 classification system is used for germline substitution classification.

Here, I describe the SBS52 classification system and how the SBS96 classification system is transformed into the SBS52 classification system. The need for the SBS52 classification system arises from the fact that certain germline substitutions are indistinguishable from one another because the reference base cannot be assumed to be the ancestral allele; as the reference genome is sequenced and assembled from a randomly sampled individual, the haplotype containing the germline mutation could have also been the

reference sequence. For instance, a C>A substitution in the AAA trinucleotide sequence context on the forward strand cannot be distinguished from a T>G substitution in the TTT trinucleotide sequence context on the reverse strand. Similarly, C>T substitutions cannot be differentiated from T>C substitutions. In addition, a C>G (T>A) substitution in a certain trinucleotide sequence context is interchangeable with another C>G (T>A) substitution in a different trinucleotide sequence context. Organised in Table 3.1 is the complete transformation of the SBS96 classification system into the SBS52 classification system.

SBS52	forward strand (reference centred)	reverse strand (read centred)
A[C>A]A	ACA>AAA = A[C>A]A	TTT>TGT = T[T>G]T
A[C>A]C	ACC>AAC = A[C>A]C	GTT>GGT = G[T>G]T
A[C>A]G	ACG>AAG = A[C>A]G	CTT>CGT = C[T>G]T
A[C>A]T	ACT>AAT = A[C>A]T	ATT>AGT = A[T>G]T
C[C>A]A	CCA>CAA = C[C>A]A	TTG>TGG = T[T>G]G
C[C>A]C	CCC>CAC = C[C>A]C	GTG>GGG = G[T>G]G
C[C>A]G	CCG>CAG = C[C>A]G	CTG>CGG = C[T>G]G
C[C>A]T	CCT>CAT = C[C>A]T	ATG>AGG = A[T>G]G
G[C>A]A	GCA>GAA = G[C>A]A	TTC>TGC = T[T>G]C
G[C>A]C	GCC>GAC = G[C>A]C	GTC>GGC = G[T>G]C
G[C>A]G	GCG>GAG = G[C>A]G	CTC>CGC = C[T>G]C
G[C>A]T	GCT>GAT = G[C>A]T	ATC>AGC = A[T>G]C
T[C>A]A	TCA>TAA = T[C>A]A	TTA>TGA = T[T>G]A
T[C>A]C	TCC>TAC = T[C>A]C	GTA>GGA = G[T>G]A
T[C>A]G	TCG>TAG = T[C>A]G	CTA>CGA = C[T>G]A
T[C>A]T	TCT>TAT = T[C>A]T	ATA>AGA = A[T>G]A
A[C>T]A	ACA>ATA = A[C>T]A	TAT>TGT = A[T>C]A
A[C>T]C	ACC>ATC = A[C>T]C	GAT>GGT = A[T>C]C
A[C>T]G	ACG>ATG = A[C>T]G	CAT>CGT = A[T>C]G
A[C>T]T	ACT>ATT = A[C>T]T	AAT>AGT = A[T>C]T
C[C>T]A	CCA>CTA = C[C>T]A	TAG>TGG = C[T>C]A
C[C>T]C	CCC>CTC = C[C>T]C	GAG>GGG = C[T>C]C
C[C>T]G	CCG>CTG = C[C>T]G	CAG>CGG = C[T>C]G

C[C>T]T	CCT>CTT = C[C>T]T	AAG>AGG = C[T>C]T
G[C>T]A	GCA>GTA = G[C>T]A	TAC>TGC = G[T>C]A
G[C>T]C	GCC>GTC = G[C>T]C	GAC>GGC = G[T>C]C
G[C>T]G	GCG>GTG = G[C>T]G	CAC>CGC = G[T>C]G
G[C>T]T	GCT>GTT = G[C>T]T	AAC>AGC = G[T>C]T
T[C>T]A	TCA>TTA = T[C>T]A	TAA>TGA = T[T>C]A
T[C>T]C	TCC>TTC = T[C>T]C	GAA>GGA = T[T>C]C
T[C>T]G	TCG>TTG = T[C>T]G	CAA>CGA = T[T>C]G
T[C>T]T	TCT>TTT = T[C>T]T	AAA>AGA = T[T>C]T
A[C>G]A	ACA>AGA = A[C>G]A	TCT>TGT = T[C>G]T
A[C>G]C	ACC>AGC = A[C>G]C	GCT>GGT = G[C>G]T
A[C>G]G	ACG>AGG = A[C>G]G	CCT>CGT = C[C>G]T
A[C>G]T	ACT>AGT = A[C>G]T	ACT>AGT = A[C>G]T
C[C>G]A	CCA>CGA = C[C>G]A	TCG>TGG = T[C>G]G
C[C>G]C	CCC>CGC = C[C>G]C	GCG>GGG = G[C>G]G
C[C>G]G	CCG>CGG = C[C>G]G	CCG>CGG = C[C>G]G
G[C>G]A	GCA>GGA = G[C>G]A	TCC>TGC = T[C>G]C
G[C>G]C	GCC>GGC = G[C>G]C	GCC>GGC = G[C>G]C
T[C>G]A	TCA>TGA = T[C>G]A	TCA>TGA = T[C>G]A
A[T>A]A	ATA>AAA = A[T>A]A	TTT>TAT = T[T>A]T
A[T>A]C	ATC>AAC = A[T>A]C	GTT>GAT = G[T>A]T
A[T>A]G	ATG>AAG = A[T>A]G	CTT>CAT = C[T>A]T
A[T>A]T	ATT>AAT = A[T>A]T	ATT>AAT = A[T>A]T
C[T>A]A	CTA>CAA = C[T>A]A	TTG>TAG = T[T>A]G
C[T>A]C	CTC>CAC = C[T>A]C	GTG>GAG = G[T>A]G
C[T>A]G	CTG>CAG = C[T>A]G	CTG>CAG = C[T>A]G
G[T>A]A	GTA>GAA = G[T>A]A	TTC>TAC = T[T>A]C
G[T>A]C	GTC>GAC = G[T>A]C	GTC>GAC = G[T>A]C
T[T>A]A	TTA>TAA = T[T>A]A	TTA>TAA = T[T>A]A

Table 3.1 SBS52 classification and corresponding SBS96 classification

After categorising germline and somatic substitutions based on the SBS52 and SBS96 classification system, SBS52 and SBS96 counts were processed as described below for *de novo* mutational signature extraction using HDP []

3.4.4.1 SBS96 mutational signature extraction

As detailed in chapter 2, SBS96 counts were normalised according to the number of callable CCS bases, callable reference bases and the trinucleotide distribution in the autosomes of the reference genome. Somatic SBS96 counts in each species is a linear combination of true positive substitutions from somatic mutational processes and false positive substitutions from library, sequencing, and software errors.

The number of false positive substitutions for each SBS96 classification, however, can be estimated from the substitution and trinucleotide sequence context dependent CCS error rate, which was determined in chapter 2, and the number of CCS trinucleotides from which somatic mutations could have been potentially called. These estimates can then be subtracted to obtain SBS96 counts where true positive substitution counts is better presented. If the same CCS library preparation protocol was not used for the DToL and the cord blood granulocyte sample, the estimation and subtraction of false positive substitutions from each SBS96 category may not be as effective in improving the signal-to-noise ratio. If a different protocol, for example, was used to extract HMW DNA and prepare a CCS library, false positive substitutions could be generated from an uncharacterized error process distinct from that identified in chapter 2.

After speciation, ongoing somatic mutational process(es) in germline stem cells depletes the trinucleotide sequence context it acts upon and shapes the trinucleotide distribution. To account for differences in trinucleotide distribution in each species, SBS96 counts are further normalised such that each trinucleotide sequence context equally contributes to the total SBS96 count (eq **??**). This normalisation further increases the signal-to-noise ratio of somatic mutational processes, particularly those with a higher somatic mutation rate and that are shared between the somatic and germline cells.

Before *de novo* mutational signature extraction, normalised SBS96 counts from each species are organised into a single matrix, samples with less than 100 somatic mutations are removed from the matrix and the total somatic mutation count for each species is normalised to the median somatic mutation count. After mutational signature extraction, each mutational signature was inspected for the following qualities to distinguish mutational signatures arising from ongoing somatic mutational processes in the sample or from library and sequencing errors upstream of sequence analysis.

1. The mutational signature is similar to those found in the COSMIC mutational signature database.

2. The mutational signature is present in another species in the same phyla

3. The mutational signature has biological replicates (e.g. idPlaAlba and xgPhoLine).
4. The mutational signature has transcriptional-strand bias
5. The mutational signature is similar to the germline mutational spectrum.
6. The attribution of the mutational signature to the total mutation burden in the sample
7. The number of somatic mutations associated with the mutational signature is not a multiple of the number of germline mutations.

3.4.4.2 Independent biological replication of mutational signatures

To confirm that identified mutational signatures are the result of a biological process and not stochastic errors,

3.4.4.3 Mutational signatures with transcriptional-strand bias

3.4.4.4 SBS52 mutational signature extraction

To *de novo* extract mutational signatures from germline mutations, germline and normalised somatic SBS52 counts from 518 eukaryotic species were organised into a single matrix and each SBS52 count was further normalised such that each trinucleotide sequence context contributes equally to the SBS52 count (eq. **??**). In contrast to the SBS96 classification system, the SBS52 classification system has 26 trinucleotide sequence contexts where the middle base is a pyrimidine base.

Through mutational signature extraction from normalised germline and somatic SBS52 counts and downstream mutational signature analysis, ancestral alleles of germline mutations were recovered and the contribution of somatic mutational processes to the germline mutational spectrum was also measured.

3.4.5 Timing the emergence of somatic mutational processes

The phylogenetic relationship between 518 species and the time of speciation was inferred from phylogenetic tree available at http://www.timetree.org and relevant information from academic literature. The birth of new species with new somatic mutational processes was used to time the emergence of new somatic mutational processes. The time at which the new somatic mutational process is estimated to have emerged will have to be updated with the ongoing efforts from the DToL consortium.

Chapter 4

Discussion

4.1 Summary of findings

In chapter 2, I hypothesise that CCS bases are, in fact, the most accurate among commercially available sequencing platforms, and I develop a tool, himut, to leverage CCS read length and base accuracy for single molecule somatic mutation detection. I benchmark himut's performance using samples where each sample has a distinct somatic mutational process and where a single somatic mutational process is responsible for newly acquired somatic mutations. The introduction of himut enables researchers to call somatic mutations, in addition to germline mutation and base modification detection from a single human genome using a single SMRTcell and the Revio sequencing instrument.

In chapter 3, I use CCS reads and high-quality reference from a range of eukaryotic species from the DToL project to study somatic mutagenesis across the tree of life. Until recently, our understanding of somatic mutational processes has been limited to species where high-quality reference genomes are available such as *H. sapiens* and model organisms such as *C. elegans* [116] and *M. musculus* [117].

To confirm that himut is applicable in non-human samples, I called somatic mutations and calculated the mutation burden of *P. lineatus* samples of various ages (3, 5 and 15). The mutation burden per cell increased with age in a clock-like fashion at a rate of 40 substitutions per cell per year, demonstrating that himut is applicable in non-human samples.

Thereafter, I use himut to detect somatic mutations in approximately 600 eukaryotic species from the DToL project. I discovered XX number of mutational signatures where SBS1-like mutational signatures were previously reported through mutational signature

analysis of somatic mutations in cancers [16] and where X number of mutational signatures (SBSX1,SBSX2, SBSX3) were new mutational signatures with an unknown aetiology.

Our analysis suggests that the emergence of a new somatic mutational process is an episodic event and once established, these processes are often conserved across species. The ubiquitous presence of the SBS1 mutational signature across the animal (annelid, bird, fish and mammal), fungi and plant (dicot) kingdom, at the earliest branching of the eukaryotic phylogenetic tree, suggests that cytosine methylation in CG dinucleotide sequence context and subsequent deamination of 5mC to thymine is an ancient process that dates back to the last eukaryotic common ancestor (LECA) or that this somatic mutational process has evolved independently in multiple eukaryotic lineages. The fact that 5mC occurs at CG, CHG and CHH sequence context in plant kingdom where H can be A, C or T nucleotides [118] and that 5mC occurs in CG dinucleotides in both animal and fungi kingdom [119] corroborates the former theory.

Germline mutation is the product of somatic mutations in germ cells and inheritance of these *de novo* mutations from one generation to the next before and after speciation. As the ancestral allele of germline mutation cannot be determined without sequence alignment with outgroup species, germline mutational processes often cannot be determined without *de novo* mutation detection through trio-sequencing. I, however, was able to use the newly extracted somatic mutational signatures to determine the germline mutational process in each species and the relative contribution of each germline mutational process in shaping the sequence context. In addition, the high similarity between the germline and somatic mutational processes suggests that the observed somatic mutational processes are clock-like mutational processes where the mutation burden increases with the age of the sample.

4.2 Limitations

4.2.1 CCS library, sequencing and software errors

CCS library preparation, sequencing and consensus sequence generation algorithm is currently not optimised to produce CCS reads where CCS bases are assigned the correct base-specific error probabilities. I limited the analysis to Q93 CCS bases as library errors and sequencing errors are unlikely to create substitution errors on both strands of the DNA and for these errors to be propagated to all the subreads during SMRT sequencing. The experimental design, hence, restricts the analysis of errors to cases where error

probabilities of Q93 CCS bases are inaccurate or to rare cases where the combination of library and sequencing errors are pervasive in CCS reads and underlying subreads.

In chapter 2, I assess Q93 CCS base accuracy using CCS reads from normal cord blood granulocytes where few somatic mutations are present. I empirically estimate that Q93 CCS base substitution error rate ranges from Q60 to Q90 depending on the substitution and the trinucleotide sequence context. In addition, I show that false positive substitutions are derived from inaccurate base accuracy estimates. What deserves the most attention is that accurate ~Q90 CCS bases can be produced for all trinucleotide sequence contexts if there are enough subreads and if correct error probabilities for subread bases are used for consensus sequence generation.

In chapter 3, I observed a somatic mutational spectrum from several species where 1) the number of called somatic mutations were greater than that from germline mutations, 2) the somatic mutational spectrum was noticeably different from the germline mutational spectrum, and finally 3) the somatic mutation spectrum was shared between phylogenetically unrelated species. In this PhD thesis, I do not investigate the origin of this somatic mutational spectrum or the downstream consequences to germline mutation detection and to assembly quality, but I hypothesise library errors to be the primary source of this erroneous somatic mutational spectrum as CCS library preparation is the only common factor in all the samples exhibiting this issue.

4.2.2 Single-molecule somatic mutation detection

In contrast to single-molecule resolution somatic mutation detection using duplex reads from the nanorate sequencing protocol, himut cannot ascertain at all trinucleotide sequence contexts whether an individual substitution, where a single read supports the mismatch between the read and the reference genome, is an error or a somatic mutation. In a clinical setting, where himut might be used to detect the earliest transformation of a normal tissue to a neoplastic tissue or to monitor tumour regression and relapse after treatment, the accuracy of every somatic mutation call is critical in helping the clinician arrive at the correct clinical interpretation. Any false-positive or false-negative mutation call could have serious consequences for the patient's treatment and prognosis.

Multiple mutational processes act on the genome at any given time and they can generate the same sequence-context specific mutation. Given a catalogue of somatic mutations from multiple samples, mutational signature analysis identifies the mutational signature in each sample and the contribution of each mutational signature to the muta-

tion burden of the sample (eq. 4.1). Each mutational signature represents the probability that a specific somatic mutational process will produce a somatic mutation in a specific sequence context.

$$M \approx PE$$
$$\begin{bmatrix} m_1^1 & \dots & m_j^1 \\ \vdots & \ddots & \vdots \\ m_1^{96} & \dots & m_j^{96} \end{bmatrix} \approx \begin{bmatrix} p_1^1 & \dots & p_s^1 \\ \vdots & \ddots & \vdots \\ p_1^{96} & \dots & p_s^{96} \end{bmatrix} \times \begin{bmatrix} e_1^1 & \dots & e_j^s \\ \vdots & \ddots & \vdots \\ e_1^s & \dots & e_j^s \end{bmatrix} \tag{4.1}$$

where M is the somatic mutation catalogue matrix with mutation type as rows and samples as columns. P is the mutational signature matrix with mutation types as rows and signatures as columns. E is the exposure matrix with signatures as rows and samples as columns. Here, I use the mutation types as defined by the SBS96 classification system for illustration purposes.

If the mutational processes and associated mutational signatures in the genome of interest are known, it is possible to calculate the probability that a given somatic mutation m in sample j originates signature s can be estimated (eq. 4.2).

$$P(m,s) = \frac{p_s^i \times e_j^s}{\sum_{s=1}^{n} p_s^i \times e_j^s} \tag{4.2}$$

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 [120]. Until the generation of error-free CCS bases, these posterior probabilities can serve as a measure of confidence for individual somatic mutations where single molecule somatic mutations are called from bulk normal tissue.

4.3 Discussion

4.3.1 Public health

The All of Us (AoU) research program aims to sequence the genomes and collect electronic health records (EHR) data of at least one million individuals in the United States from under-represented demographic categories to accelerate biomedical research [121]. The

AoU research program can use himut to investigate somatic mutagenesis across all their samples where CCS reads are available, just as I have used himut to study somatic mutational processes across the tree of life. The sheer number of samples sequenced under the AoU research program will enable the discovery of new mutational signatures resulting from environmental mutagenesis, DNA damage and mismatch repair deficiencies, as well as their possible combinations. Moreover, AoU research program can also leverage the EHR records (e.g. age of the sample, geographical location, dietary and drinking habits and drug prescription history) to develop and evaluate hypotheses about the aetiology of the newly discovered mutational signatures.

The tobacco smoking mutational signature (SBS4) is a canonical example of a mutational signature where exogenous exposure to a mutagen (tobacco carcinogen) is responsible for somatic mutagenesis. The elevation of mutation burden attributable to SBS4 in smokers compared to non-smokers suggests that lung cancer, linked to tobacco smoking, is a preventable disease [122]. Aristolochic acid (AA) consumption, often through traditional Chinese medicine, is an under-recognised source of somatic mutagenesis (SBS22) and is a major contributor to endemic Balkan nephropathy [123] and urinary tract urothelial carcinoma in Taiwan [124]. The discovery of somatic mutagenesis resulting from inadvertent or involuntary exposure to carcinogens, hence, might be one of the most intriguing outcomes of population-scale CCS sequencing efforts.

4.3.2 DNA forensics

DNA fingerprinting is often used in criminal investigations to determine whether the reference DNA sample from the crime scene and DNA sample from the suspect is derived from the same individual [125]. If the genetic sequence of two random individuals are compared, 99.9% of their genetic sequences are estimated to be identical [27]. The number of core repeat motifs in variable number tandem repeat (VNTR) loci, however, is unique to each individual and DNA fingerprinting leverages variation in VNTR loci as a unique genetic fingerprint to compare and match DNA samples from different individuals.

The age of the sample is another unique biomarker that can facilitate the identification of an individual. Cells accrue somatic mutations in a clock-like fashion. Haematopoietic stem cells, for example, acquire 16.8 substitutions per cell per year [114, 101] while sperm cells with the lowest somatic mutation rate accumulate 2.9 substitutions per cell per year [126]. In addition to the ability to call somatic mutations, himut can also calculate the mutation burden per cell. The age of the sample at the time it was collected, therefore, can

be derived from mutation burden of the sample and the tissue-specific somatic mutation rate. The age of the sample in question, thereafter, can either help investigators narrow the number of suspects or free innocent individuals.

4.3.3 The birth and death of somatic mutational processes

Every living species on Earth is thought to be the direct descendent of the last universal common ancestor (LUCA). The tree of life symbolises how different species share a common ancestor and how they have diverged over time through speciation and natural selection. Somatic mutagenesis in parental cells of asexually reproducing species or gametes of sexually reproducing species is at the heart of evolution as somatic mutational processes generate the genetic diversity that natural and sexual selection act upon for adaptation and speciation. Using the CCS reads and high-quality reference genomes from the DToL project, I had the privilege to detect, analyse and discover new and conserved somatic mutational processes across the tree of life and time the emergence of these somatic mutational processes in evolutionary time scale.

As discussed in chapter 3, in many eukaryotic species, the same somatic mutational process is responsible for generating mutations in both gametes and somatic cells, but in a subset of species such as the *P. albimanus* (white-footed hoverfly) and *S. pipiens* (thicked-legged hoverfly) somatic mutational process of an unknown aetiology is not only distinct from the germline mutational processes, but also exhibits transcriptional-strand bias. The presence of the same somatic mutational process in biological replicates confirms that the observed mutational spectrum is a consequence of a specific biological process. The strength of these somatic mutational processes relative to the background mutational spectrum perhaps suggests that they might be resulting from endogenous or exogenous stress.

The presence of SBS1 mutational signature in animal (annelids, birds, fish, mammals), fungi and plant kingdom, the three eukaryotic kingdoms that represents the earliest diversification of the eukaryotic lineage, underscores the importance of cytosine methylation as an epigenetic mechanism. In *C. elegans*, the loss of DNA methyltransferase 1 (DNMT1), a drastic measure, is required to eliminate cytosine methylation and subsequent deamination of 5mC to thymine.

In the light of this fact, the absence of SBS1 mutational signature in the insect phylum is another striking discovery that highlights the taxonomic differences in somatic mutational processes. In addition, the dearth of C>T somatic mutations in honeybees that use

cytosine methylation for caste differentiation is another observation that corroborates this phenomenon [127]. Considering the importance of cytosine methylation as an epigenetic modification, I hypothesise that either the insect phylum has evolved a DNA damage repair pathway that is more efficient than that operational in other phyla or another base modification has been selected as the primary epigenetic modification.

The discovery of a greater number of unique somatic mutational processes in the insect phylum, compared to other phyla, is another unexpected revelation. The insect phylum boasts the greatest diversity among all phyla in the animal kingdom not only in the number of individuals, but also the number of species. The greater diversity of somatic mutational processes is presumably then linked to the successful adaptation and speciation of insect phylum in ecological niches where other phyla have been unsuccessful. The absence of additional somatic mutational processes that act upon the gametes in other phyla of the animal kingdom suggests that other somatic mutational processes have a selective disadvantage to the individual. If this is indeed true, a natural question arises: why are new somatic mutational processes harmful to individuals of non-insect phyla, and what properties have allowed SBS1 mutational signature to be conserved across so many species for billions of years? Furthermore, how has the insect phylum evolved not just one new somatic mutational process, but multiple new processes? And how has the DNA damage pathway evolved to accommodate the dramatic changes in somatic mutational process that occur during speciation?

The DToL project currently focuses on sequencing eukaryotic species in the UK and Ireland. As the genomes of all living species and the next generation of species are a cumulative result of somatic mutagenesis, the expansion of the DToL project to include archaea and bacteria, the other two domains of life, could potentially bring us closer to unravelling how life started on Earth. To answer the question 'what is life?', origin of life chemists have used a bottom-approach to synthesise organic molecules that constitutes life and mimic the emergent properties of life [128]. In parallel, synthetic biologists have used a top-down approach to create a minimal viable cell [129]. The identification of somatic mutational processes across all domains of life, determination of ongoing somatic mutational processes in each species and the accurate delineation of phylogenetic relationship between species could potentially help us design an alternative top-down approach to model the possible nucleotide composition of LUCA and derive the sequence events of that happened at the start of life.

4.3.4 Evolutionary advantage of complete metamorphosis

Coleoptera (beetles), hymenoptera (sawflies, bees, wasps, and ants), diptera (midges, mosquitoes, and flies) and lepidoptra (moths and butterflies), insects that undergo complete metamorphosis, account for more than 80% known insect species and 95% of total insect species diversity (Fig 4.1). Many theories have been proposed to explain the evolutionary advantage of complete metamorphosis such as the decoupling of growth and differentiation [130], the ability to exploit different environments [131], and reduction in competition between juvenile insect and adult insect for limited resources [132]. Here, I hypothesise that the complete metamorphosis impedes the transmission of somatic mutations acquired during the juvenile stage to the adult insect.

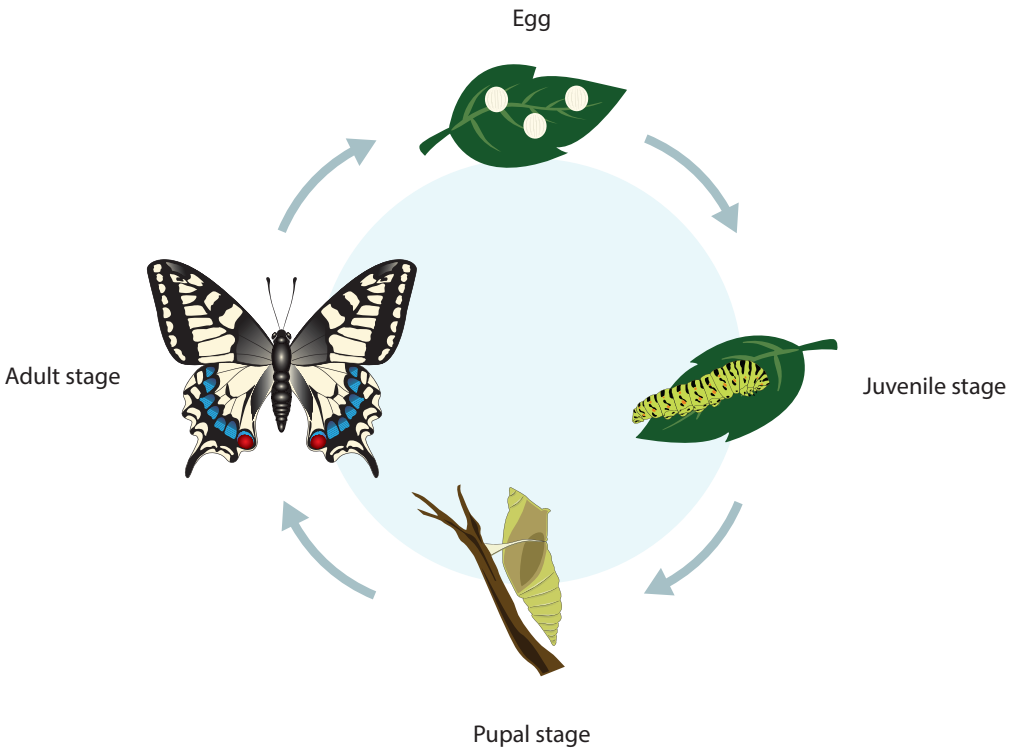


Fig. 4.1 The life cycle of a *Papilo machaon*

The somatic mutation theory of ageing suggests that the gradual accumulation of somatic mutations leads to a decline in cellular function, which ultimately contributes to the ageing process [133]. The theory also implies that shorter-lived species will have a higher somatic mutation rate than longer-lived species. Since somatic mutation rate is inversely proportional to the lifespan of species in mammals [134], I initially conjectured

that somatic mutation rate would be higher in insects as most insects are short-lived [135]. I observed that mutation burden of insects was lower than what might have been expected from the somatic mutation theory of ageing and an alternative explanation was required for this observation (The DToL project sequences one sample per species and age of the sample is often not recorded. The calculation of somatic mutation rate requires multiple samples of different ages).

As a disposable tissue that does not contribute to the germ line lineage, the human placenta has a higher mutation burden and chromosomal aberrations that are absent from the foetus [51]. On the other hand, the human spermatogonia has the lowest somatic mutation rate [126]. Similarly, larval tissue in holometabola (a superorder of insects that undergoes complete metamorphosis) is a disposable tissue like the placenta that does not contribute to the development of the adult insect, and hence, larval tissue does not transmit genetic information to the next generation. During embryonic development, imaginal disc precursors separate from embryonic stem cells programmed for differentiation into larval tissue after the blastoderm stage and imaginal discs are set aside for further development into the structures of the adult insect during the pupal stage (An imaginal disc is a single layer of epithelial sheet that consists of 20 to 40 cells) [136]. Genitalia, for example, are derived from the medial disc in *D. melanogaster* [137]. Larva and adult insects are, therefore, derived from distinct embryonic lineage, aside from the histoblast cells that develop into the abdomen of the adult insect [137]. Consequently, somatic mutations that were acquired in the larval stage should be absent in the adult insect and only the somatic mutations that arose during the first few cell divisions of embryonic development should be shared between the two tissues (Fig 4.2). For instance, caterpillars, the larvae of butterflies and moths, primarily consume the leaves, stems and flowers of plants and the subsequent accumulation of chlorophyll pigment might be phototoxic to caterpillars. During photosynthesis, the photoexcitation of chlorophyll pigment is channelled to convert light energy into glucose and oxygen. In contrast, unregulated photoexcitation of chlorophyll pigment produces reactive oxygen species [138]. Hence, C>A somatic mutations associated with DNA damage by reactive oxygen species (SBS18 and SBS36) might the dominant somatic mutational process in caterpillars.

The life cycle and development of holometabolous insects corroborates the hypothesis that somatic mutations acquired during the larval stage are not shared with the adult insect. In addition, the hypothesis could be tested with insect species that exhibit sexual dimorphism in development. In certain insect species, females do not undergo metamorphosis, referred to as larviform female, while the males still undergo complete

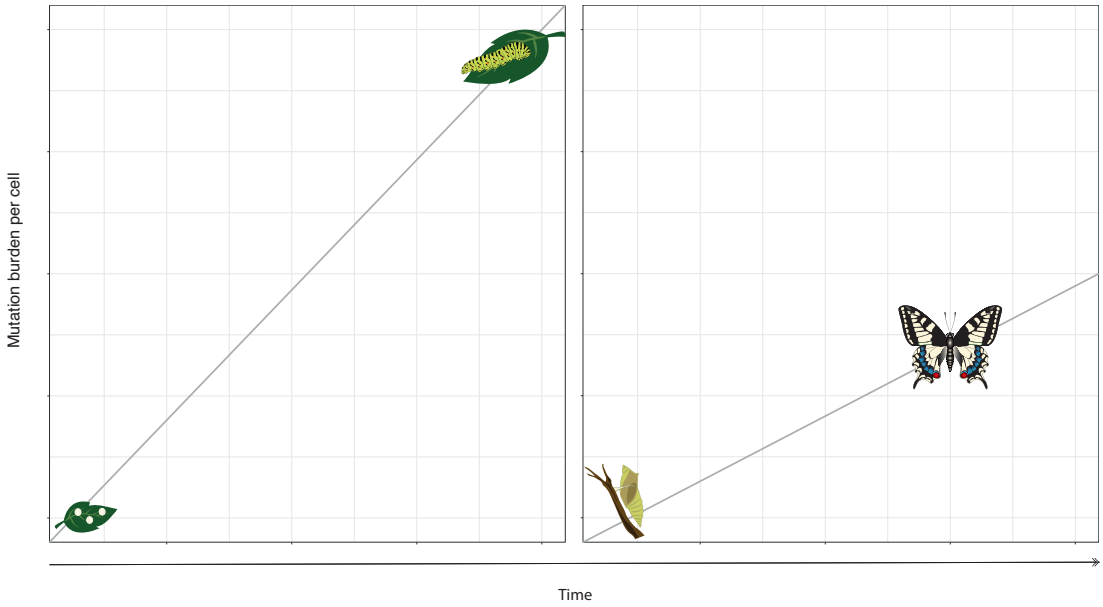


Fig. 4.2 Hypothetical changes in mutation burden with life cycle progression

Larval tissue is hypothesised to have a higher somatic mutation rate than adult tissue. Somatic mutations acquired during the juvenile stage are not transmitted to the adult insect and a new somatic mutational process is operational in the adult tissue.

metamorphosis. CCS sequencing of both male and female samples with such sexual dimorphism could confirm whether somatic mutation rate is higher in the larval tissue. This experiment, however, does not address the question of why the mutation burden is low in adult insects despite the high cell division rate during metamorphosis.

4.4 Future directions

In the imminent future, I conjecture that CCS library errors and inaccurate CCS BQ score estimation will be properly addressed and that the majority (>50%) of CCS bases will have ~Q90 base accuracy. Here, I discuss the potential opportunities following this development.

4.4.1 Single-molecule real-time sequencing

As discussed in chapter 1, CCS sequence throughput and sequencing cost per base is a function of the number of ZMWs and the read-of-insert length of the SMRTbell template. As demonstrated in chapter 2 and chapter 3, CCS base accuracy is a function of subread error rate and the number of subreads per CCS read, under the assumption that there are no new errors introduced during consensus sequence generation. Here, I make some informed predictions about the forthcoming advancements in the SMRT sequencing platform based on observations made in this PhD thesis.

I expect the number of ZMWs per SMRTcell to double every two to three years like how Moore's Law predicts the number of transistors per chip to double every two years. As the number of ZMWs per SMRTcell increases exponentially, sequencing cost per base is expected to decrease exponentially as well (Fig 4.3a). Moore's law has continued for approximately ~50 years and similar performance increases can be expected from SMRTcell as well. In addition, as CCS sequence throughput is directly proportional to CCS read length, the rate at which CCS sequence throughput increases could also exceed all our expectations due to the combined effect of parallel increases in both CCS read length and number of ZMWS per SMRTcell (Fig 4.3b).

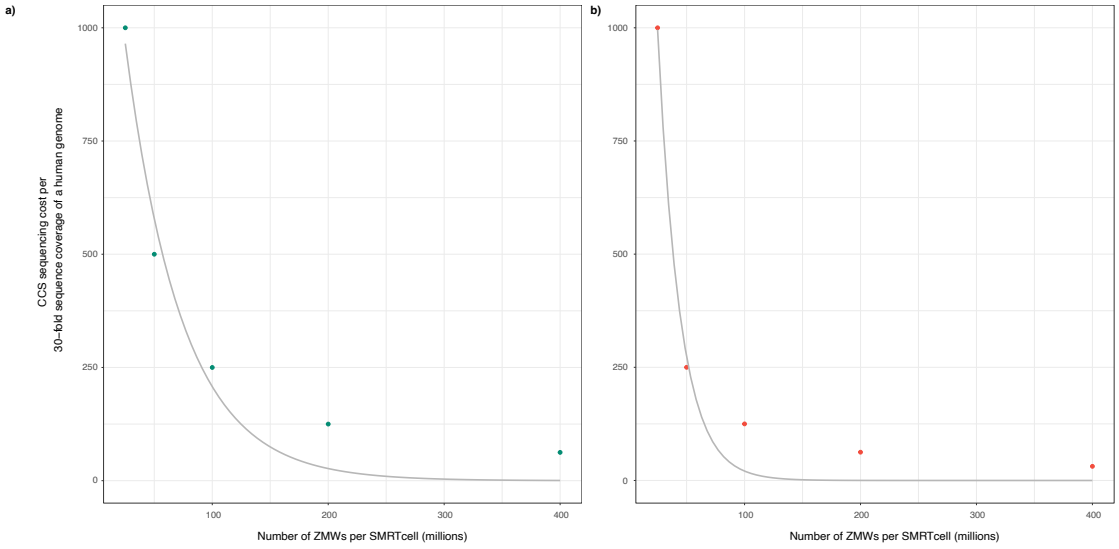


Fig. 4.3 Exponential decay in CCS sequencing cost

The graph starts with the current CCS sequencing cost for a 30-fold sequence coverage of a human genome using the Revio instrument, the latest SMRTcell with 25 million ZMWs and assumes that average CCS read length is around 20kb. **a)** Exponential decay in CCS sequencing cost with doubling in the number of ZMWs per SMRTcell. **b)** Exponential decay in CCS sequencing cost with doubling in both CCS read length and the number of ZMWs per SMRTcell.

To make accurate predictions about future technological advances, I must also consider Wright's Law, a companion of Moore's Law. Wright's Law, also known as experience curve effect, states that for every cumulative doubling of units produced, costs will fall by a constant percentage. As discussed above, the rapid decrease in CCS sequencing costs will accelerate the adoption of CCS sequencing and when economies of scale are achieved, the positive flywheel effect will be unstoppable (Fig 4.4).

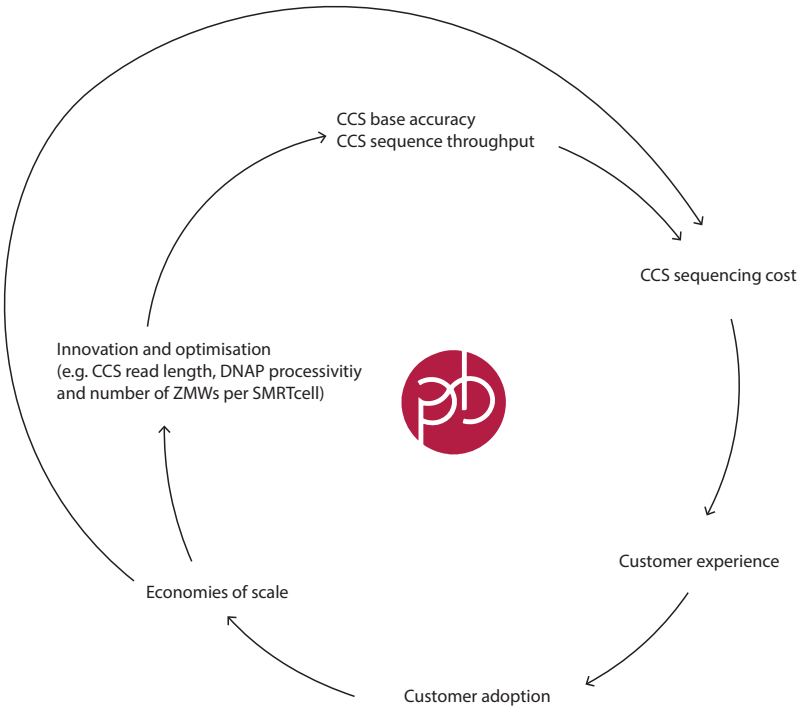


Fig. 4.4 Pacific Biosciences flywheel

The flywheel symbolises how independent components act in concert to improve base accuracy, reduce sequencing cost and drive customer adoption

DNAP processivity, the rate at which DNAP synthesises a new strand of DNA, is another crucial factor that determines CCS base accuracy and sequence throughput. If the subread error rate is between 10% and 15%, CCS generation typically requires at least 10 subreads per CCS read to generate CCS bases with Q93 base accuracy. DNAP processivity determines the number of subreads per CCS read and hence, increasing DNAP processivity translates to increasing CCS read length. If, for example, DNAP processivity is doubled, CCS read length can also be doubled without sacrifices in base accuracy. DNAP's biological limit, hence, will be the theoretical limit of CCS read length. In addition, DNAP replication

error rate is not an obstacle to ~Q90 CCS base generation at all trinucleotide sequence contexts as demonstrated in chapter 2. One interesting ramification of improved base accuracy is that pooled and non-barcoded samples can be sequenced together to detect common germline SNPs as lower sequence coverage is needed to call the germline mutations with confidence. In addition, as the number of somatic mutations increases linearly with sequence coverage, CCS sequence coverage will not determine the confidence with which germline mutations are detected, but the number of somatic mutations that are detected from the sample.

In contrast, Illumina's sequencing by synthesis approach has several disadvantages that limit improvements in read length and base accuracy. In each Illumina sequencing cycle, the rate at which a growing DNA becomes asynchronous with the rest of DNA fragments from the same cluster increases, resulting in a reduced signal-to-noise ratio as sequencing progresses [139]. This technical limitation places a ceiling on Illumina read length and is responsible for decline in per-base sequence quality towards the end of the read. In addition, CCS sequence throughput is a polynomial function with read length and number of ZMWs per SMRTcell as input while Illumina sequence throughput is a linear function with number of clusters per flow cell as the only input. Consequently, CCS sequencing possesses a greater potential for improving sequence throughput. Considering the fact that CCS reads enables *de novo* assembly and simultaneous detection of haplotype phased somatic and germline mutations and epigenetic modifications, I believe that CCS sequencing will be the primary DNA sequencing method in clinics and research in the imminent future.

4.4.2 Strand-specific somatic mutation detection

To date, somatic mutation detection in normal tissues and tumours with next-generation sequencing have focused on analysing sub-clonal or clonal somatic mutations that are fixed in a group of cells above the limit of detection threshold. Single-molecule resolution and strand-specific base modification somatic mutation detection has the potential to enhance our understanding of somatic mutagenesis.

As described in chapter 1, DNAP sequences both the forward and reverse strand of the SMRTbell template multiple times through rolling circle replication. The pbccs algorithm leverages the redundancies and complementary base pairing between the forward and reverse strand subreads to generate CCS reads. As demonstrated in chapter 2, CCS reads have sufficient base accuracy for single molecule somatic mutation detection and as

described in a previous publication, single-molecule resolution 5mC detection is also possible from CCS DNAP kinetics [140, 141]. The pbccs algorithm can also generate single-strand consensus sequence (SSCS) reads from the forward and reverse strand subreads. Strand-specific somatic mutation and base modification detection with SSCS reads presents an exciting opportunity to analyse somatic mutations from their beginning to their end.

Somatic mutation is a three-step process: 1) DNA damage, mutation, or modification from endogenous or exogenous sources, 2) failure to detect and repair the DNA damage or mutation, and 3) fixation, persistence of DNA mutation in daughter cells through genetic drift or selection. In a population of DNA molecules, there will be a group of wild type DNA molecules, a group of DNA molecules with DNA damage, mutation or modification, a group of DNA molecules undergoing DNA damage repair and a group of DNA molecules with new somatic mutations (Fig 4.5).

Fig. 4.5 SBS1 somatic mutational process

The figure illustrates how DNA mismatch resulting from spontaneous deamination of 5mC to thymine is repaired and how DNA mutation is fixed. The nucleotide bases in red highlights the methylated CG dinucleotide on both the forward and reverse strand of the DNA.

The DNA damage and repair process associated with SBS1 mutational signature, for example, is amenable to further qualitative and quantitative examination through this approach. The spontaneous deamination of 5mC to thymine results in a TG:GC mismatch and results in C>T somatic mutation at a CG dinucleotide if left unrepaired by the mismatch repair (MMR) pathway. If both strands of the double-stranded DNA molecule are sequenced, TG dinucleotide will be present on the strand where deamination has happened and GC dinucleotide with methylation will be present on the complementary strand. SSCS reads from SMRT sequencing enable the detection of TG:GC mismatches and associated hemi-methylation (Figure 4.5). In addition, CCS reads allow the estimation of the number of methylated CG dinucleotides where deamination could have happened and the number of CG dinucleotides where somatic mutations have occurred. If the same tissue is sequenced at multiple different timepoints, the gain and loss of somatic mutations in the population can also be studied.

If successful, we will be able to measure the *in vivo* deamination rate from the number of TG:GC mismatch and the number of GC dinucleotides, and compare it against the *in vitro* deamination rate of 5.8×10^{-13} per 5mC per second at 37°C [142]. In addition, TG:GC mismatch repair efficiency and fidelity can also be measured under wild-type and mutant conditions. MutSα, for example, is critical in recognising the TG:GC mismatch and initiating DNA damage repair. MutSα deficiency, therefore, elevates the number of C>T somatic mutations. Similarly, cross-examination of both SSCS and CCS reads and associated DNAP kinetics can also be used to better understand the C>T (SBS2) somatic mutations resulting from APOBEC-dependent deamination of cytosine to uracil.

4.4.3 Decomposition of a mutational signature

Single-molecule resolution and strand-specific base modification somatic mutation detection, most importantly, creates an opportunity to gain greater insights into the dynamics of somatic mutational processes. Each somatic mutational process leaves a characteristic imprint to the genome and mutational signatures represent the probability that a specific somatic mutational process will produce a somatic mutation in a specific sequence context. Given a catalogue of somatic mutations from multiple samples, mutational signature analysis identifies the mutational signature in each sample and the contribution of each mutational signature to the mutation burden of the sample.

Since each mutational signature is a cumulative result of DNA damage, mutation or modification, failure to repair the DNA damage or mismatch, and persistence of the mutation in bulk tissue, each mutational signature can be re-defined as such (eq. 4.3)

$$\begin{gathered} \alpha D \cdot \beta R \cdot \gamma F \approx P_i \\ \alpha \begin{bmatrix} d_1^1 \\ \vdots \\ d_1^{96} \end{bmatrix} \beta \begin{bmatrix} r_1^1 \\ \vdots \\ r_1^{96} \end{bmatrix} \gamma \begin{bmatrix} f_1^1 \\ \vdots \\ f_1^{96} \end{bmatrix} \approx \begin{bmatrix} p_1^1 \\ \vdots \\ p_1^{96} \end{bmatrix} \end{gathered} \tag{4.3}$$

where D is the DNA damage matrix and each element is a probability that a specific sequence context will be damaged. R is DNA damage repair matrix and each element is a probability that DNA damage in a specific sequence context will be repaired. F is DNA mutation fixation matrix and each element is a probability that the mutation type will be fixed in the population. α, β, γ are scalar values that represent genetic and environmental factors that modulate the somatic mutational process. In addition, R could be further

decomposed into multiple subcomponents where each matrix represents a different DNA damage repair pathway specific to the DNA damage (eq. 4.4)

$$\alpha \begin{bmatrix} d_1^1 \\ \vdots \\ d_1^{96} \end{bmatrix} \beta \left[\begin{bmatrix} r_i^1 \\ \vdots \\ r_i^{96} \end{bmatrix} \begin{bmatrix} r_{i+1}^1 \\ \vdots \\ r_{i+1}^{96} \end{bmatrix} \dots \begin{bmatrix} r_n^1 \\ \vdots \\ r_n^{96} \end{bmatrix} \right] \gamma \begin{bmatrix} f_1^1 \\ \vdots \\ f_1^{96} \end{bmatrix} \approx \begin{bmatrix} p_1^1 \\ \vdots \\ p_1^{96} \end{bmatrix} \tag{4.4}$$

The decomposition of a mutational signature into their individual components and subcomponents should enable us to have a greater understanding of the nonlinear relationship between the components and the mechanisms underlying somatic mutagenesis.

4.4.4 Gene conversion and crossover detection

Here, I also hypothesise that CCS read length and base accuracy can be leveraged to detect gene conversion and crossover events generated during meiotic and mitotic recombination. Gene conversion and crossover (CO) (Fig 4.6) arise from the non-reciprocal and reciprocal exchange of genetic material during double-strand break (DSB) repair through homologous recombination. Gene conversions are also referred to as non-crossovers (NCO), but gene conversions and crossovers are not mutually exclusive events [143].

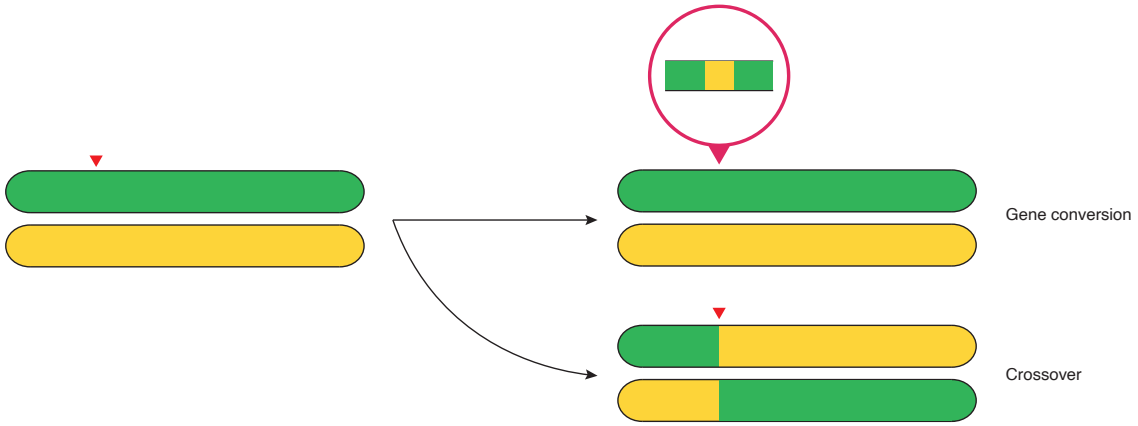


Fig. 4.6 Gene conversion and crossover

Red triangle indicates the DSB site. DSB can be repaired either as a gene conversion or as a crossover. Sister chromatids are not shown for simplification purposes.

In germ cells, meiotic recombination is an essential process that generates new combinations of alleles that serve as the foundation for adaptation and speciation through natural selection, an advantage for sexually reproducing organisms. In addition, the formation of at least one chiasma per pair of homologous chromosomes ensures proper segregation of chromosomes in anaphase I of meiosis. Improper chromosome segregation can result in aneuploid gametes with abnormal numbers of chromosomes. If DSB repair is not repaired in somatic cells, DNA damage response initiates programmed cell death. It is worth noting that DSB repair during meiotic recombination generates new allele combinations and contributes to genetic diversity, while DSB repair during mitotic recombination can result in the loss of heterozygosity (LOH). Furthermore, meiotic DSBs are deliberately introduced through the concerted action of PRDM9 and SPO11 to initiate meiotic recombination while mitotic DSBs are inadvertently generated from both endogenous (e.g. reactive oxygen species) and exogenous factors (e.g. ionising radiation).

A haplotype is a set of alleles that are inherited together in a single chromosome. Meiotic recombination yields a new haplotype with a new combination of alleles. Haplotype phasing, therefore, is not only essential for determining the original haplotype before meiotic recombination, but also the haplotype phase switch in individual gametes or in children after meiotic recombination. Trio-sequencing [144], sperm-typing [145] and statistical methods leveraging the non-random association of alleles (linkage disequilibrium) [146] have been previously used to detect meiotic recombination products and each of these approaches have trade-offs in the resolution and the number of meiotic recombination event detected. Trio-sequencing approach, for example, detects one meiotic recombination per chromosome per child and enables the study of sex-specific meiotic recombination rate. In contrast, sperm-typing of a bulk sperm sample can detect multiple meiotic recombination events per target locus, but it requires prior knowledge of meiotic recombination hotspots. In addition, while statistical methods can generate high-resolution maps of recombination events from patterns of LD in the human population, they are unable to examine meiotic recombination events in an individual.

CCS sequencing of a bulk sperm sample should enable unbiased and genome-wide detection of meiotic recombination events. In addition, the number of detected events should be directly proportional to the sequence coverage as one chiasma per pair of homologous chromosomes is required for physical pairing and proper segregation of chromosomes. If a locus is a meiotic recombination hotspot, CCS sequencing of bulk sperm sample will yield three population of CCS reads: those derived from maternal haplotype, those derived from the paternal haplotype and those resulting from meiotic

recombination and containing both maternal and paternal haplotypes. As discussed in chapter 2, CCS read length and base accuracy can be leveraged to phase hetSNPs and connect phased SNPs to construct a haplotype block. CCS haplotype, subsequently, can be determined from comparing CCS hetSNPs to that of the haplotype block. Similarly, CCS reads with both maternal and paternal haplotype can be identified from such a comparison. The length, and the number of phase transitions will also inform whether holliday junction (HJ) was resolved as a gene conversion or a crossover. It is worth noting that the resolution of the recombinant product detection and the number of loci from which recombinant products can be called from is higher in samples with greater heterozygosity. As detailed in chapter 3, many eukaryotic samples have a higher density of SNPs than human samples. In short, CCS base accuracy is used to not detect single molecule somatic mutations but detect single molecule phase transitions (Fig 4.7).

Our approach also has several advantages compared to existing methods and can greatly contribute to the body of research attempting to understand the hotspot conversion paradox [147]. *De novo* mutations, for example, can be detected using himut and mutation burden across multiple samples of different ages can be used to define the germline mutation rate at scale in species where germline mutation rate is unknown. The hotspot conversion paradox stems from the fact that GC biased gene conversion leads to the loss of PRDM9 recognition sites and meiotic recombination hotspots. PRDM9 gene binds to a specific DNA sequence motif (CCNCCNTNNCCNC) [148] and initiates meiotic recombination through the recruitment of SPO11 for programmed induction of double-strand breaks. The polymorphisms in the zinc finger array determine the exact DNA sequence motif and the PRDM9 binding site. The ability to determine the PRDM9 allele, identify PRDM9 *de novo* mutations and detect gene conversion events using CCS reads should enable us to better understand how the rapid evolution of the PRDM9 gene resolves this paradox.

The application of the above approach to somatic cells should enable the comprehensive characterisation of mitotic recombination and an assessment of how loss of heterozygosity contributes to oncogenesis in normal cells. Bloom syndrome patients, for example, with an autosomal recessive mutation in the BLM gene have increased risk of developing cancer [149]. In addition, cross-examination of both meiotic and mitotic recombination products should elucidate the similarities and differences between these two processes and enhance our understanding of speciation and tumorigenesis.

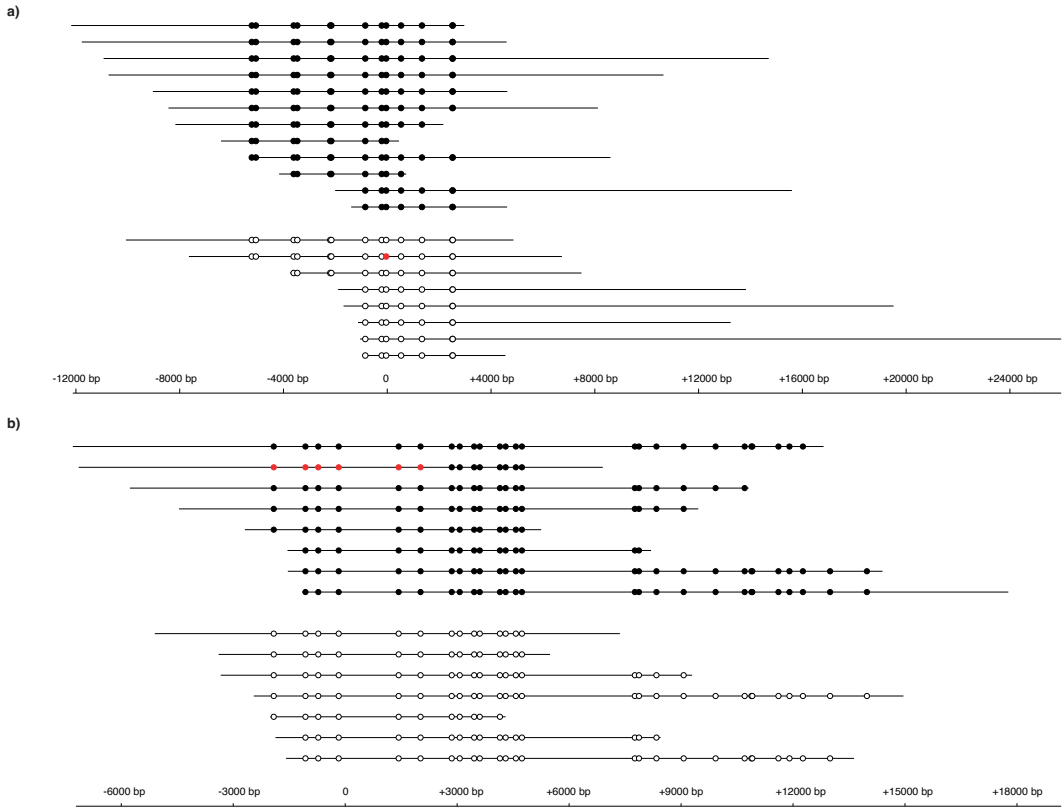


Fig. 4.7 Gene conversion and crossover detection using CCS reads

Each circle indicates a heterozygous SNP. A black circle indicates a reference allele, white circle indicates an alternative allele and red circle indicates a phase transition from reference allele to alternative allele and vice versa. CCS reads derived from the same haplotype have the same set of heterozygous SNPs. **a)** Gene conversion detection using CCS reads requires the phase transition to be flanked by wild type alleles of the haplotype. **b)** Crossover detection using CCS reads necessitates the phase transition to be continuous.

4.5 Concluding remarks

I conclude that factors that prevent the adoption of CCS sequencing are technical problems where solutions exist. As discussed in chapter 2, almost error-free CCS bases can be generated and as conjectured in this chapter, CCS sequencing cost and HMW DNA input requirement for CCS library preparation will no longer be a limitation to research. The exponential increase in the number of ZMWs per SMRTcell and the read-of-insert length will be the primary factors driving the increase in sequence throughput and decrease in per-base sequencing cost. The present sequencing methods necessitates a specific DNA input requirement to sequence the genome multiple times and thereby enable the detec-

tion of mutations with greater confidence despite the presence of sequencing errors. If DNAP processivity improves to enable CCS library preparation from longer read-of-insert and if CCS base accuracy improves to be error-free, only a single read will be required from each haplotype for germline and somatic mutation detection and epigenetic modification identification, drastically lowering the HMW DNA input requirements for CCS library preparation. I believe that we are witnessing a historic moment where error-free sequencing will be feasible at a fraction of current sequencing costs and where it will be possible to interrogate the genetic, epigenetic, and transcriptomic information of all forms of life.

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] 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.

[110] 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.

[111] pysam developers. pysam.

[112] 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.

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

[114] 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.

[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.

[116] Bettina Meier, Susanna L Cooke, Joerg Weiss, Aymeric P Bailly, Ludmil B Alexandrov, John Marshall, Keiran Raine, Mark Maddison, Elizabeth Anderson, Michael R Stratton, Anton Gartner, and Peter J Campbell. C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res.*, 24(10):1624–1636, October 2014.

[117] Laura Riva, Arun R Pandiri, Yun Rose Li, Alastair Droop, James Hewinson, Michael A Quail, Vivek Iyer, Rebecca Shepherd, Ronald A Herbert, Peter J Campbell, Robert C Sills, Ludmil B Alexandrov, Allan Balmain, and David J Adams. The mutational signature profile of known and suspected human carcinogens in mice. *Nat. Genet.*, 52(11):1189–1197, November 2020.

[118] Ian R Henderson and Steven E Jacobsen. Epigenetic inheritance in plants. *Nature*, 447(7143):418–424, May 2007.

[119] Adam J Bewick, Brigitte T Hofmeister, Rob A Powers, Stephen J Mondo, Igor V Grigoriev, Timothy Y James, Jason E Stajich, and Robert J Schmitz. Diversity of cytosine methylation across the fungal tree of life. *Nat Ecol Evol*, 3(3):479–490, March 2019.

[120] Eric Letouzé, Jayendra Shinde, Victor Renault, Gabrielle Couchy, Jean-Frédéric Blanc, Emmanuel Tubacher, Quentin Bayard, Delphine Bacq, Vincent Meyer, Jérémy Semhoun, Paulette Bioulac-Sage, Sophie Prévôt, Daniel Azoulay, Valérie Paradis, Sandrine Imbeaud, Jean-François Deleuze, and Jessica Zucman-Rossi. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.*, 8(1):1315, November 2017.

[121] The "all of us" research program.pdf.

[122] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J Campbell, Paolo Vineis, David H Phillips, and Michael R Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, November 2016.

[123] Arthur P Grollman, Shinya Shibutani, Masaaki Moriya, Frederick Miller, Lin Wu, Ute Moll, Naomi Suzuki, Andrea Fernandes, Thomas Rosenquist, Zvonimir Medverec, Krunoslav Jakovina, Branko Brdar, Neda Slade, Robert J Turesky, Angela K Goodenough, Robert Rieger, Mato Vukelić, and Bojan Jelaković. Aristolochic acid and the etiology of endemic (balkan) nephropathy. *Proc. Natl. Acad. Sci. U. S. A.*, 104(29):12129–12134, July 2007.

[124] Chung-Hsin Chen, Kathleen G Dickman, Masaaki Moriya, Jiri Zavadil, Viktoriya S Sidorenko, Karen L Edwards, Dmitri V Gnatenko, Lin Wu, Robert J Turesky, Xue-Ru Wu, Yeong-Shiau Pu, and Arthur P Grollman. Aristolochic acid-associated urothelial cancer in taiwan. *Proc. Natl. Acad. Sci. U. S. A.*, 109(21):8241–8246, May 2012.

[125] Peter Gill, Alec J Jeffreyst, and David J Werrett. Forensic applications of dna 'fingerprints'.

[126] Raheleh Rahbari, Arthur Wuster, Sarah J Lindsay, Robert J Hardwick, Ludmil B Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, Michael R Stratton, UK10K Consortium, and Matthew E Hurles. Timing, rates and spectra of human germline mutation. *Nat. Genet.*, 48(2):126–133, February 2016.

[127] G Luo, I M Santoro, L D McDaniel, I Nishijima, M Mills, H Youssoufian, H Vogel, R A Schultz, and A Bradley. Cancer predisposition caused by elevated mitotic recombination in bloom mice. *Nat. Genet.*, 26(4):424–429, December 2000.

[128] A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, May 1953.

[129] Clyde A Hutchison, 3rd, Ray-Yuan Chuang, Vladimir N Noskov, Nacyra Assad-Garcia, Thomas J Deerinck, Mark H Ellisman, John Gill, Krishna Kannan, Bogumil J Karas, Li Ma, James F Pelletier, Zhi-Qing Qi, R Alexander Richter, Elizabeth A Strychalski, Lijie Sun, Yo Suzuki, Billyana Tsvetanova, Kim S Wise, Hamilton O Smith, John I Glass, Chuck Merryman, Daniel G Gibson, and J Craig Venter. Design and synthesis of a minimal bacterial genome. *Science*, 351(6280):aad6253, March 2016.

[130] Jens Rolff, Paul R Johnston, and Stuart Reynolds. Complete metamorphosis of insects. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 374(1783):20190063, October 2019.

[131] Charles Darwin. *On the Origin of Species by Means of Natural Selection.* John Murray, London, 1859.

[132] Bo Ebenman. Evolution in organisms that change their niches during the life cycle. *Am. Nat.*, 139(5):990–1021, May 1992.

[133] L Szilard. On the nature of the aging process. *Proc. Natl. Acad. Sci. U. S. A.*, 45(1):30–45, January 1959.

[134] Alex Cagan, Adrian Baez-Ortega, Natalia Brzozowska, Federico Abascal, Tim H H Coorens, Mathijs A Sanders, Andrew R J Lawson, Luke M R Harvey, Shriram Bhosle, David Jones, Raul E Alcantara, Timothy M Butler, Yvette Hooks, Kirsty Roberts, Elizabeth Anderson, Sharna Lunn, Edmund Flach, Simon Spiro, Inez Januszczak,

Ethan Wrigglesworth, Hannah Jenkins, Tilly Dallas, Nic Masters, Matthew W Perkins, Robert Deaville, Megan Druce, Ruzhica Bogeska, Michael D Milsom, Björn Neumann, Frank Gorman, Fernando Constantino-Casas, Laura Peachey, Diana Bochynska, Ewan St John Smith, Moritz Gerstung, Peter J Campbell, Elizabeth P Murchison, Michael R Stratton, and Iñigo Martincorena. Somatic mutation rates scale with lifespan across mammals. *Nature*, 604(7906):517–524, April 2022.

[135] Daniel E L Promislow, Thomas Flatt, and Russell Bonduriansky. The biology of aging in insects: From drosophila to other insects and back. *Annu. Rev. Entomol.*, 67:83–103, January 2022.

[136] Clonal analysis of primordial disc cells in the early embryo of drosophila melanogaster.

[137] Michael Bate and Martinez Alfonso Arias. *The Development of Drosophila melanogaster*. Cold Spring Harbor Laboratory Press, New York, 1993.

[138] Christine H Foyer. Reactive oxygen species, oxidative signaling and the regulation of photosynthesis. *Environ. Exp. Bot.*, 154:134–142, October 2018.

[139] Michael L Metzker. Emerging technologies in dna sequencing. *Genome Res.*, 15(12):1767–1776, December 2005.

[140] Joaquim S L Vong, Peiyong Jiang, Suk-Hang Cheng, Wing-Shan Lee, Jason C H Tsang, Tak-Yeung Leung, K C Allen Chan, Rossa W K Chiu, and Y M Dennis Lo. Enrichment of fetal and maternal long cell-free DNA fragments from maternal plasma following DNA repair. *Prenat. Diagn.*, 39(2):88–99, January 2019.

[141] O Y Olivia Tse, Peiyong Jiang, Suk Hang Cheng, Wenlei Peng, Huimin Shang, John Wong, Stephen L Chan, Liona C Y Poon, Tak Y Leung, K C Allen Chan, Rossa W K Chiu, and Y M Dennis Lo. Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 118(5), February 2021.

[142] J C Shen, W M Rideout, 3rd, and P A Jones. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded dna. *Nucleic Acids Res.*, 22(6):972–976, March 1994.

[143] Neil Hunter. Meiotic recombination: The essence of heredity. *Cold Spring Harb. Perspect. Biol.*, 7(12), October 2015.

[144] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, Sigurjon A Gudjonsson, Michael L Frigge, Agnar Helgason, Unnur Thorsteinsdottir, and Kari Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature*, 467(7319):1099–1103, October 2010.

[145] Adam J Webb, Ingrid L Berg, and Alec Jeffreys. Sperm cross-over activity in regions of the human genome showing extreme breakdown of marker association. *Proc. Natl. Acad. Sci. U. S. A.*, 105(30):10471–10476, July 2008.

[146] Simon Myers, Leonardo Bottolo, Colin Freeman, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. *Science*, 310(5746):321–324, October 2005.

[147] A Boulton, R S Myers, and R J Redfield. The hotspot conversion paradox and the evolution of meiotic recombination. *Proc. Natl. Acad. Sci. U. S. A.*, 94(15):8058–8063, July 1997.

[148] Simon Myers, Colin Freeman, Adam Auton, Peter Donnelly, and Gil McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, 40(9):1124–1129, September 2008.

[149] Stephen B Gruber, Nathan A Ellis, Karen K Scott, Ronit Almog, Prema Kolachana, Joseph D Bonner, Tomas Kirchhoff, Lynn P Tomsho, Khedoudja Nafa, Heather Pierce, Marcelo Low, Jaya Satagopan, Hedy Rennert, Helen Huang, Joel K Greenson, Joanna Groden, Beth Rapaport, Jinru Shia, Stephen Johnson, Peter K Gregersen, Curtis C Harris, Jeff Boyd, Gad Rennert, and Kenneth Offit. BLM heterozygosity and the risk of colorectal cancer. *Science*, 297(5589):2013, September 2002.