Chapter 1

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

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

1.1 The Genomics Revolution

1.1.1 Genome assembly

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

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

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

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

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

1.1.2 Human Genome Project

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

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

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

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

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

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

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

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

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

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

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

How is physical mapping done? Contamination removal

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

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

1.1.3 ~~Illumina sequencing~~

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

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

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

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

Somatic mutation detection, however, is not a solved problem. Somatic mutation callers, for example, employ different strategies and exhibit varying specificities and sensitivities. Consensus somatic mutation call from multiple different detection algorithms, hence, is often used for downstream analysis [20]. The base accuracy and read length of Illumina reads, most importantly, is the common technical factor that limits the resolution at which the somatic mutations can be detected. MuTect, for example, cannot differentiate Illumina sequencing errors from low variant allele fraction (VAF) somatic mutations as a typical Illumina base call has a 0.01-1% error rate [21]. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations [22–24]. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure **??**). The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells.

The repeat content of the genome is another hurdle for accurate somatic mutation detection. Repetitive sequences (e.g., tandem repeat expansions, retrotransposons, segmental duplications, telomeric repeats and centromeric alpha-satellite) account for approximately 50% of the human genome [25]. If the repeat length is greater than the read length, read alignment software cannot determine the location of the read with respect to the reference genome as the read could have originated from any copies of the repetitive sequence [26]. The accurate placement of reads, hence, requires repetitive sequences to be flanked with unique sequences not present elsewhere in the reference genome. Consequently, the reference genome is divided into callable regions 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, improves the accuracy and precision of both read alignment and variant calling [33].

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

1.1.4 Challenges in somatic mutation detection

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

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

Somatic mutation detection, however, is not a solved problem. Somatic mutation callers, for example, employ different strategies and exhibit varying specificities and sensitivities. Consensus somatic mutation call from multiple different detection algorithms, hence, is often used for downstream analysis [20]. The base accuracy and read length of Illumina reads, most importantly, is the common technical factor that limits the resolution at which the somatic mutations can be detected. MuTect, for example, cannot differentiate Illumina sequencing errors from low variant allele fraction (VAF) somatic mutations as a typical Illumina base call has a 0.01-1% error rate [21]. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations [22–24]. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure **??**). The detection of these somatic mutations, however, are critical for early detection of cancer, monitoring of tumour evolution during patient treatment and to enhance our understanding of the transformation of normal cells to neoplastic cells.

1.1.5 Single molecule somatic mutation detection

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

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

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

1.2 Single molecule sequencing

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

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

1.2.1 Nanopore sequencing

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

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

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

1.2.2 Pacific Biosciences Single-Molecule Real-Time sequencing

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

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

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

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

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

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

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

1.2.3 Long-read sequencing applications

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

1.2.3.1 *De novo* assembly

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

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

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

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

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

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

1.2.3.2 Full-length transcript sequencing

In contrast to short-read sequencing that requires *de novo* assembly of RNA reads to acquire full-length transcripts, long-read sequencing can be used to obtain full-length transcripts 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 offer compared to short-read sequencing technologies for somatic structural rearrangement detection, the application of long-read sequencing technologies to somatic mutation detection has been limited to date. There has been a handful publications that interrogated somatic structural rearrangements in breast cancer cell lines with long reads []. Somatic mutation detection with long reads is at the stage where we are re-creating the capabilities provided by short-sequencing technology and is not at the stage where we are finding somatic mutations that cannot be detected with short-read sequencing technology.

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

1.3 The Darwin Tree of Life Project

1.4 Origins of Life

1.4.1 Prebiotic Earth

1.4.2 The Garden of Eden

1.4.3 RNA world

1.4.4 The emergence and evolution of life on Earth

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

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

1.5 Thesis objectives

The history of science is ~~riddled with~~ full of 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.