**\section{The Genomics Revolution}**

**\subsection{Sanger dideoxy sequencing: the beginning}**

The definitive identification of DNA as the genetic that material is responsible for transformation of R strain of bacteria into S strain bacteria by Oswald Avery in 1944\cite{} and the discovery of the structure of the DNA by James Watson and Francis Crick in 1953\cite{} has sparked interest to develop DNA sequencing technologies to determine the nucleotide composition of organisms and to better understand genetic variations. In 1977, Frederick Sanger developed chain termination method\cite{}, and Allan Maxam and Walter Gilbert developed chemical cleavage method\cite{} for DNA sequencing. The chain termination method, also referred to as Sanger dideoxy sequencing, became the preferred method for DNA sequencing thanks to scalable implementation of the method using fluorescent nucleotides by Applied Biosystems (ABI) \cite{} and was the primary sequencing method for the Human Genome Project.

To obtain a higher confidence for each Sanger read base calls, the complementary strand can also be sequenced\cite{} Sanger sequencing is the original duplex sequencing method (discussed later).

The original chain termination method starts with a single-stranded template DNA and a primer designed to bind to the start of the single-stranded template DNA. DNA polymerase (DNAP) binds to the primer and initiates DNA elongation using the free nucleotides in the mixture. Sanger mixed used both deoxyribonucleotides and dideoxyribonucleotides where the concentration of deoxyribonucleotides were higher than the concentration of dideoxyribonucleotides such that DNA elongation will be preferred. Until the incorporation of the dideoxyribonucleotide, DNAP will use the template DNA and perform DNA elgonation. Upon the incorporation of dideoxyribonucleotide, DNA elongation is terminated as the dideoxyribonucleotide does not have the 3’-OH to form phosphodiester bond with next nucleotide. Repeat chain-termination experiments results in DNA fragments of varying sizes and these fragments are ordered by their size through gel electrophoresis. This sequencing experiment is repeated with the four dideoxynucleotides (ddATP, ddGTP, ddCTP and ddTTP) such that the DNA sequence can be determined from reading the gel image from top to bottom \cite{}. Sanger and colleagues used this method to determine the sequence of the 5,375 bp long $\Phi$X174 bacteriophage. Sanger sequencing at the time produced Sanger reads with ~200bp in length and reads with overlapping 5’ end and 3’ end were manually inspected and connected to obtain the $\Phi$X174 bacteriophage genome.

Upon the completion and refinement of the Sanger sequencing method, the race began to sequence the smallest genome, the entire genetic information of an organism, and to progressively sequence and assemble larger and more complex genomes\cite{}

In addition, the need for a software that can find overlaps between Sanger reads, to inspect the overlaps and to connect the overlapping reads into a single contiguous sequence, referred to as a contig, became apparent\cite{}. In addition, to distinguish overlaps produced as a result sequencing errors from true overlaps, uncertainty associated with each base was calculated.

Hood and colleagues modified the Sanger sequencing method to use fluorescently labelled nucleotides and demonstrated that the emitted fluorescence from the chain terminating dideoxynucleotide can be used to determine the nucleotide sequence.

ABI modified the chain-termination method such that fluorophore-labelled dideoxynucleotide is used and this allowed sequencing from reading the fluorescence emitted from the chain-terminating nucleotide. The first iteration of the high-throughput chain-termination method produces reads in 500bp length, and the use of Taq polymerase with higher processivity increased the read length to up to 1000 bp long.

*Following rounds of template DNA extension from the bound primer, the resulting DNA fragments are heat*[*denatured*](https://en.wikipedia.org/wiki/DNA_denaturation)*and separated by size using*[*gel electrophoresis*](https://en.wikipedia.org/wiki/Gel_electrophoresis)*.*

*The DNA sample is divided into four separate sequencing reactions, containing all four of the standard*[*deoxynucleotides*](https://en.wikipedia.org/wiki/Deoxynucleotides-triphosphate)*(dATP, dGTP, dCTP and dTTP) and the DNA polymerase.  To each reaction is added only one of the four*[*dideoxynucleotides*](https://en.wikipedia.org/wiki/Dideoxynucleotides)*(ddATP, ddGTP, ddCTP, or ddTTP), while the other added nucleotides are ordinary ones. The deoxynucleotide concentration should be approximately 100-fold higher than that of the corresponding dideoxynucleotide (e.g. 0.5mM dTTP : 0.005mM ddTTP) to allow enough fragments to be produced while still transcribing the complete sequence (but the concentration of ddNTP also depends on the desired length of sequence).*[*[4]*](https://en.wikipedia.org/wiki/Sanger_sequencing#cite_note-Sanger1977-4)

**During sample preparation, different-sized fragments of DNA are generated each starting from the same location.**

**Each fragment ends with a particular base that is labeled with one of four fluorescent dyes corresponding to that particular base. Then all of the fragments are distributed in the order of their length via capillary electrophoresis. Information regarding the last base is used to determine the original sequence.**

**\subsection{Genome assembly}**

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

The ability to determine the nucleotide composition of organisms at scale with ABI capillary sequencing platform initiated a race to determine the genome sequence of scientific and economic interest and to determine the method that is most suitable for the human genome project. In principle, genome assembly aims to use randomly selected DNA fragments from the genome, to find overlaps between the DNA fragments and to connect the overlaps into a single contiguous sequence. If the genome in question does not have repeats or if the read length is greater than repeat length, genome assembly becomes a trivial problem. Repeats account for less than X\% of prokaryotic genomes. Repeats, however, are common in eukaryotic genomes and account for ~50\% of the human genome. Repeats take many forms and repeats can exist as tandem repeats, palindromes, or inverted repeats. There are repeats created by retrotransposons where retrotransposons use copy and paste mechanisms to create copies of themselves in the genome. Segmental duplications is a special type of repeat where non-repetitive sequences greater than 1kb with interchromosmal or intrachromosomal duplications with sequence identity greater than 99% \cite{}. 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\cite{}

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.

Read and error rate.

**\subsection{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 \cite{}, 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 \cite{}.

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 \cite{}. Shotgun sequencing aims to assemble the genome from random DNA fragments sampled from the genome. Simulations has shown that if paired-end sequencing is performed on inserts of vary length with sufficient coverage, sufficient overlaps can be found to create contigs. In addition, mate-pairs can, thereafter, be used to order and orient contigs into scaffolds. Shotgun sequencing was proposed as an alternative to hierarchical shotgun sequencing approach as shotgun sequencing approach would not require the creation of BAC clones libraries, physical mapping of the BAC clones and independent sequencing and assembly of the BAC clones, thereby reducing the cost of the genome assembly drastically.

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

In contrast, NIH preferred hierarchical shotgun sequencing, also known as clone-by-clone, approach for construction of the human reference genome as the aims of NIH was not to create the assembly in shortest time, but to create a reference genome that can withstand the test of time and that can act as a focal point for scientific research and for scientific community. The hierarchical shotgun sequencing approach simplifies the assembly problem to the assembly of the 50-100kb BAC clone. Upon the successful assembly of the BAC clone, the location of the BAC contig can be determined from physical maps and overlapping BAC contigs can be assembled into a unitig \cite{}. 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 \cite{}. 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 \cite{}. 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 \cite{}.

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 \cite{}.

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\cite{}. Segmental duplications have resulted in human-specific gene duplications not found in other great apes \cite{}, 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 \cite{}.

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 \cite{}.

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 \cite{}, 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 \cite{}.

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

**\section{Next-generation sequencing revolution}**

\textit{Any sufficiently advanced technology is indistinguishable from magic}

\begin{flushright} [Arthur C. Clarke] \end{flushright}

The 20th century had been an amazing decade of series of discoveries in biological sciences. Erwin Schrodinger proposed in his famous lecture “What is Life” in 1943 that physical material responsible for genetic inheritance must have an aperiodic crystal structure. Since the identification of DNA as the transforming agent of R strain to S strain by Oswald Avery \cite{} and the discovery of the structure of the DNA by James Watson and Francis Crick in 1954 \cite{}, there is now no question that DNA is the physical material responsible for genetic inheritance and the genome, the entire corpus of genetic information of an organism, dictates the embryonic development, cell differentiation and function, organisation of tissues, and even the life span of an organism.

Today, sequencing instruments function as a modern microscope not to look at physiology of individual or population of cells, but to measure and interpret genetic variations (DNA) and their phenotypic manifestations (RNA and amino acids) to understand the intricacies of cellular functions. Ion-torrent \cite{}, Pyro \cite{}, Roche 454 \cite{} and Illumina \cite{} sequencing platforms once competed in the sequencing market as next-generation sequencing platforms and Illumina platform has emerged as the sequencing method of choice such that Illumina platform has become synonymous with the next-generation sequencing. The advent of high-throughput sequencing instrument combined with exponential decrease in sequencing costs has completely transformed how we approach biological questions.

**\subsection{Illumina Platform}**

Illumina platform performs sequencing-by-synthesis (SBS) to generate paired-end reads, each of 150bp read length. A typical read from the Illumina platform is shorter than that produced from Sanger sequencing, but the shorter read length is compensated with redundant sequencing of the genome.

***contact, detect, dissociate, de-block, incorporate, rinse, repea***

%% sequencing cost, moore’s law

%% economics of scale

%% innovation on top of next-generation sequencing

%% ancient genomics

%% iteration of illumine sequencing machines: GAII

%% hi-c sequencing: loops configurations, topologically associating domain (TADs), A/B compartments

%% chip sequencing

%% 3C sequencing (Job Dekker)

The technical limitations of Illumina sequencing (base accuracy and short read length), however, has been the bottleneck for improving rare genetic disease diagnostics yield, detecting rare somatic mutations and constructing high-quality reference genomes for non-human species. De novo assembly of other species, previously, have been attempted using de brugjin graph based de novo assembly algorithms with short reads, but assemblies produced from short reads were highly fragmented and incomplete. In addition, scaffolding strategies often did not provide sufficient long-range information to produce chromosome-level pseudomolecules and as a result, these assemblies provided incomplete information for comparative genomics purposes. Hence, assemblies produced from short reads often have collapsed repeats or contigs that cannot be placed accurately.

%% population-genetics

%% population genomics

%% increase in the number of whole-genome sequenced with illumine sequencing

%% cancer genomics, driver mutation, mutational signatures

%% clinical sequencing

%% tumour evolution

%% liquid biopsy

The human genome project is estimated to have cost 3 billion dollars, equivalent to 1 dollar per base pair. And as technology becomes more ubiquitous and democratised, we have constantly shifted/moved from studying one individual to studying the group. We initially focused on studying a single individual and as sequencing cost has decreased, population genomic studies and the history of differences and going back in time to study our lineage. What is common and different.

**\subsection{Data Standardisation}**

A wave of standardisation to create file formats that is universally accepted across the community.

The technical limitations of the Illumina platform limits the interrogation of the genome and the inability to access the “dark matter” of the genome \cite{} and to improve clinical diagnosis with Illumina platform has convinced other researchers to use other sequencing technologies.

**\subsection{Short-read sequencing applications}**

**\subsubsection{Bisulfite sequencing}**

**\subsubsection{Single-cell sequencing}**

**\subsubsection{High-throughput chromatin conformation capture sequencing}**

**\subsubsection{}**

\section{}

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 \cite{Stratton2009-of}. 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 \cite{Martinez-Jimenez2020-kn}, identify biological processes that generate these mutations \cite{Alexandrov2013-kg}, to use somatic mutations as timestamps and biological barcodes to lineage trace development \cite{Behjati2014-gb}, to discover complex structural rearrangements such as chromothripsis \cite{Stephens2011-gj} that fundamentally changed the conventional view of tumorigenesis as the gradual process of the accumulation of somatic mutations \cite{Doll1954-of, Knudson1971-fg} and to better understand the relationship between abnormal embryonic development and paediatric tumour formation \cite{Marshall2014-ec}. International efforts such as the Cancer Genome Atlas (TCGA) program \cite{Weinstein2013-ko} and the Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium \cite{ICGCTCGA\_Pan-Cancer\_Analysis\_of\_Whole\_Genomes\_Consortium2020-ts} 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 \cite{Hanahan2000-dp, Hanahan2011-zr}.

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 \textit{de novo} extract new mutational signatures or to assign the contribution of known mutational signatures to the mutation burden \cite{Alexandrov2013-fq}; 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 \cite{Alexandrov2013-fq, Li2020-vw, Steele2022-mn}.

The PCAWG consortium has discovered 67 single-base-substitution (SBS) mutational signatures \cite{Alexandrov2020-ys}. 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 \cite{Degasperi2022-qe}. In addition, somatic mutations resulting from chemotherapeutic agents is another active area of research \cite{Pich2019-ja, Aitken2020-sa}. 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 \cite{Bailey2020-ou}. 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 \cite{Cibulskis2013-gw}. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations \cite{Costello2013-cz, Chen2017-ba, Abascal2021-pk}. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure \ref{}). 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 \cite{Lander2001-du}. 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 \cite{Li2008-dt}. 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 \cite{1000\_Genomes\_Project\_Consortium2012-rj}. Clinically relevant genes in non-callable regions, hence, are often excluded from analysis \cite{Wagner2022-ph}.

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 \cite{Lander2001-du}, 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 \cite{Osoegawa2001-np}. 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 \cite{Garrison2018-ae}. The Genome Reference Consortium (GRC) has released GRCh38 build with alternative loci to address some of these issues \cite{Schneider2017-yo}. 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 \cite{Nurk2022-dv}. T2T-CHM13 genome, as expected, improve the accuracy and precision of both read alignment and variant calling \cite{Aganezov2022-dv}.

Table of current somatic mutation callers, their sensitivity and specificity, and their approaches \cite{}.

**\subsection{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 \cite{Bailey2020-ou}. 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 \cite{Cibulskis2013-gw}. Library errors, introduced upstream of sequencing, are also often misclassified as somatic mutations \cite{Costello2013-cz, Chen2017-ba, Abascal2021-pk}. Newly acquired somatic mutations, therefore, are indistinguishable from background noise using conventional methods and required breakthroughs in sample and library preparation (Figure \ref{}). 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.

**\subsection{Single molecule somatic mutation detection}**

Illumina’s technical limitations have limited somatic mutation detection to clonal or sub-clonal 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 \cite{Lodato2018-hh}, single-cell clone expansion \cite{Lee-Six2018-qe} and laser-capture microdissection (LCM) \cite{Ellis2021-it} and sequencing adopts the former approach. Rolling circle amplification \cite{Lizardi1998-qh, Dahl2004-tm} and duplex sequencing methods \cite{Schmitt2012-yr, Abascal2021-pk, Hoang2016-jx} 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 \cite{Lee-Six2018-qe, Martincorena2015-gu, Ju2017-vw, Martincorena2018-av, Brunner2019-xg, Lee-Six2019-vt, Yoshida2020-yr, Olafsson2020-vi, Moore2020-pi, Lawson2020-em, MSpencer\_Chapman2021-cq, Coorens2021-ct, Robinson2021-te, Grossmann2021-gd, Moore2021-dl, Park2021-fx, Ng2021-jd}. 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 \cite{Newman2016-cy}.