**\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" \o "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).