# BT2062 - Analytical Techniques in Biotechnology

Assignment 1 - Human Genome Sequencing

*Submitted by Sahana Gangadharan (BE17B038)*

## 1 Human Genome

The human genome is a complete set of nucleic acids encoded as the deoxyribonucleic acid (DNA) within 23 chromosomes pairs that exist in the nucleus of every cell. It is known as the nuclear genome. A smaller amount of DNA exists within the cellular organelle mitochondria, and that is known as the mitochondrial genome. Hence, the human genome comprises both nuclear and mitochondrial genomes. The human genome consists of protein-coding and protein non-coding DNA segments.

## 2 Early Sequencing - The first generation

Fred Sanger sequenced the first biomolecule fragment, in the early 1950s. Insulin was the first protein that was sequenced. Sanger fragmented the two chains of Insulin, deciphered them individually, and finally overlapped the fragments to yield the overall sequence of the protein. He mentioned that sequencing of a protein in that manner offered insights on the specific patterns of amino acid residues, found in proteins [1]. Following this vanilla technique was the method of Edman degradation, a repeated elimination of an N-terminal residue from the peptide chain [2]. This method made sequencing easier both experimentally and analytically. Although the above techniques were tedious and cumbersome, many proteins were sequenced by the late 1960s, using the above-said methods.

RNA Sequencing was also attempted using a similar technique in the 1960s. The RNA sequence would first be treated with RNases, followed by separation using chromatography and electrophoresis. The broken individual fragments were then deciphered by sequential exonuclease digestion, and the overlaps were fit into position to identify the whole sequence. The first RNA sequence, of alanine tRNA, was found out using one gram of pure material (isolated from 140 kg of yeast) to determine 76 nucleotides [3]. This entire process was majorly simplified by ‘fingerprinting’ techniques, which included the separation of radioactively labelled RNA fragments and visualization in two dimensions [4]. The positions from the visualization are indicative of the size and the sequence of the fragments.

Early attempts at DNA sequencing were also unwieldy, just like the above methods. Wu and Kaiser attempted to sequence the 12 bases of the cohesive ends of bacteriophage lambda using primer extension methods, in 1968 [5]. Following this, Gilbert and Maxam reported the 24 bases of the lactose-repressor binding site, in 1973. This was performed by copying the DNA sequence into RNA

and sequencing those fragments. By around 1976, the field experienced a sudden surge with the development of two techniques by Sanger & Coulson and Maxam & Gilbert.

## 2.1 Sanger Sequencing - Principle and Protocol

Popularly referred to as the "Sanger's Sequencing", the former method was a "*chain termination technique*" that helped in finding the distance of nucleotides from a radioactive label [6]. This method offers 99.99% accuracy in the sequence and is considered the 'gold standard' for DNA sequencing. The protocol of the technique is as follows.

The DNA sequence of interest is used as a template, in a special kind of PCR, known as the chain-termination PCR. Along with the dNTPs added into the PCR mixture, four types of dideoxynucleotide triphosphates (ddNTPs) are also added. The ddNTPs lack the 3′-OH group (they have an H attached at the 3' end) that is required to form a phosphodiester bond, which is the primary linkage that holds the nucleotides in place. Therefore, when the ddNTP is incorporated by the DNA polymerase into the strand while copying from the template, the elongation stops there and the polymerase cannot form a link with the next nucleotide. The user mixes ddNTPs and dNTPs in a lower ratio so that sequences at all lengths are formed. Each ddNTPs is further labelled with a distinct fluorescent dye, that will be useful in analysis. In manual Sanger sequencing, four PCR reactions are set up, one corresponding to each ddNTP and the four regular dNTPs. The oligonucleotides from each of the PCR reactions are run in four separate lanes in gel electrophoresis, thus allowing the user to identify the exact position of each nucleotide, moving bottom to top[1]. In both the above techniques, with one lane per identification of one base, the gels were placed under X-ray film producing a ladder image, from which the sequence could be read off immediately.

In conclusion, the Sanger sequencing consists of the following six steps [7]-

1. The double-stranded DNA (dsDNA) is denatured into two single-stranded DNA (ssDNA).
2. A primer that corresponds to one end of the sequence is attached.
3. Four polymerase solutions with four types of dNTPs are composed, but only a single type of ddNTP is added.
4. The DNA synthesis reaction initiates, and the chain extends until a termination nucleotide is randomly incorporated.
5. The resulting DNA fragments are denatured into ssDNA.
6. The denatured fragments are separated by gel electrophoresis, and the sequence is determined.

[1] **Note**: In gel electrophoresis, the sequence of larger length does not diffuse through the gel much faster compared to shorter fragments that are known to diffuse very quickly. Hence, while inferring the sequence, the user must go from bottom to top.

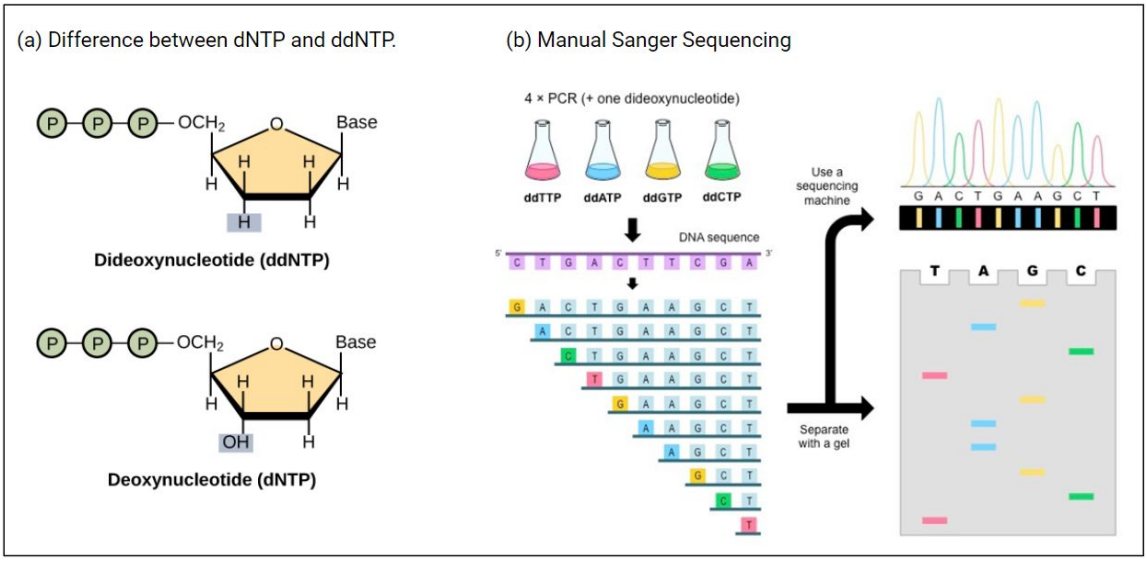


**Figure 1:** Sanger sequencing - (a) The structure of a ddNTP and a dNTP. **Source**: "Whole-genome sequencing: Figure 1," by OpenStax College, Biology (b) Manual sequencing protocol. Each lane corresponds to each of the four added ddNTPs. The sequence is read from bottom to top. **Source** - The Genome education, Facebook.

## 2.2 Gilbert and Maxam method for Sequencing

This method also involves radioactive labelling of nucleotides to identify the precise positions of the bases. They describe reactions that cleave DNA at Adenines (A), at Adenines and Guanines (A+G), at Thymines (T) and, Cytosine and Thymines (C+T). To cleave A+G, dimethyl sulfate is added. This compound methylates Gs at N7 position and As at N3 position, thus making the glycosidic bond unstable and cleaving them. Understanding that glycosidic bond of methylated adenosine is much weaker than that of guanine, a simple treatment with acid will cleave bonds at As. It has been known that Hydrazine reacts with Cytosines and Thymines to cleave the bases and the presence of 2M NaCl suppresses the reaction of thymines with hydrazines, thereby preferentially yielding oligonucleotides that end at Cytosines. This way, the nucleotides are sequentially cut at all the specified positions. After dephosphorylation, the two 5′ ends of the fragment are labelled with 32P for tracking purposes. The rest of the protocol is very similar to that of Sanger's. Purified DNA could be used directly in this method, while the Sanger method requires that each read start is cloned for production of single-stranded DNA [8].

**Interpreting the results:** The positions of A and C can be identified from bottom to top from the respective lanes. However, thymine residues are identified at positions that have a band on the C+T lane, but not in the corresponding C lane. Similarly, guanine residues are identified at positions that have a band on the A+G lane, but not in the corresponding A lane.

## 2.3 Shotgun Sequencing

The next remarkable technique that was widely used in genome sequencing is the method of Shotgun sequencing - sequencing random fragments of the DNA sequence and then aligning them based on overlapping parts to identify the complete sequence [9]. Staden proposed this technique employing three computer programs for the searching, alignment and joining of the sequences. The program OVRLAP searches for overlaps between sequences from different gel readings or clones. XMATCH is a program which allows the operator to scrutinise overlapping sequences, edit them and join them together. FILINS is a general program for the manipulation of sequence files.

By 1987, around 1000 bases per day was generated by automated fluorescence-based Sanger sequencing machines, developed by Smith, Hood and Applied Biosystems [10]. This led to the massive growth of the sequence-data, that were (are) stored in repositories such as GenBank, which further gave rise to essential search tools such as BLAST.

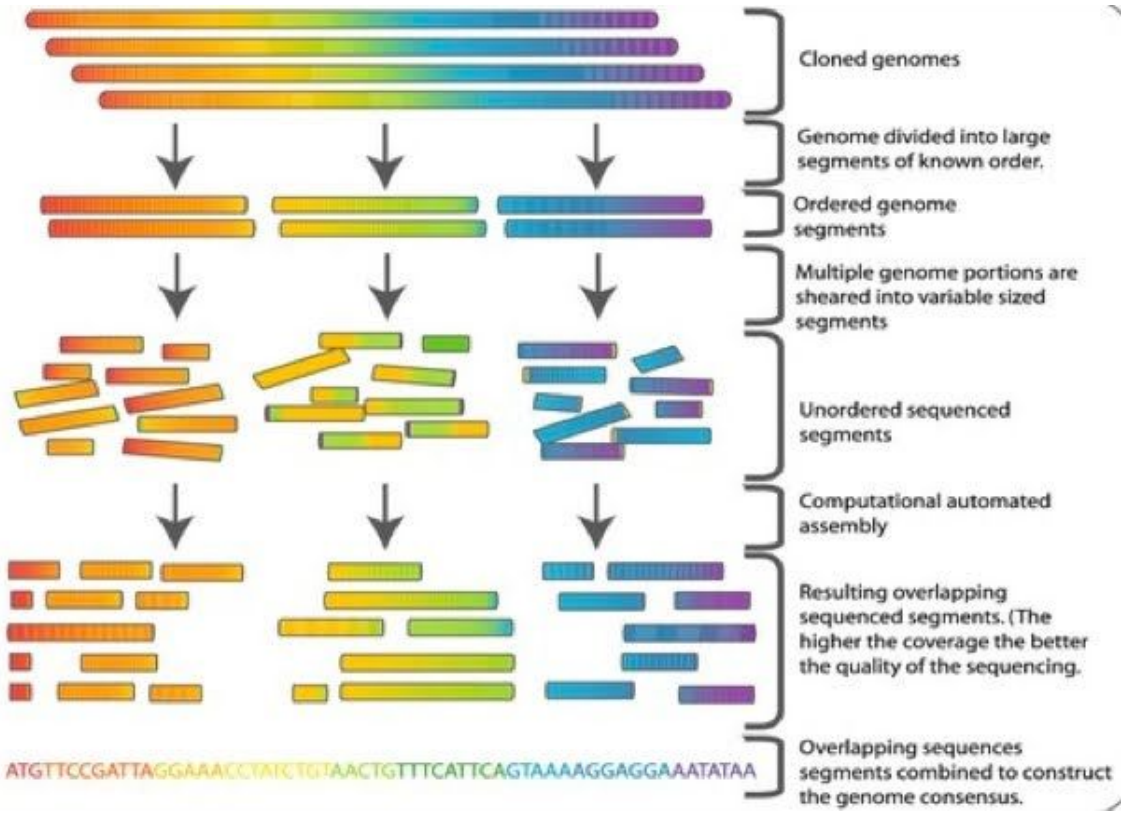


**Figure 2**: Shotgun sequencing and assembly.

# 3 The Human Genome Project (HGP)

The main objective of the HGP was to determine the DNA sequence of the entire eukaryotic human genome. The publicly funded project commenced in 1990 and was officially closed in April 2003. The genome (of total length around 3 billion bases) was broken down into large numbers of smaller fragments of length 150,000 base pairs approximately, using **endonucleases**. These known endonucleases cleave the genomic DNA at known sequences using the **Restriction Digestion** technique. Several restriction enzymes such as HindIII, EcoRI and PstI were used in this procedure. This leaves the cleaved site a blunt/sticky end for the ligation with BAC or plasmid.

Each fragment was then ligated into a vector referred to as the **Bacterial Artificial Chromosome** (**BAC**)[2]. Bacterial Artificial Chromosomes (BAC) is a large segment of DNA (100,000—200,000 bp) from another species cloned into bacteria. This creates a BAC-derived hierarchical shotgun library. Further, these vectors are cloned inside bacteria, where each fragment is increased in numbers by the DNA replication machinery.

Each library clone exhibited a DNA fragment "fingerprint," which could be compared to that of all other library clones to identify overlapping clones. This method is also known as “restriction fingerprinting” as it identifies a set of restriction sites contained in each clone. In possible areas where the sequence and the chromosomal location of DNA fragments are known, these fragments within the library vectors were mapped to their respective chromosomal regions by screening for **sequence-tagged sites (STSs),** that can be amplified using **polymerase chain reaction (PCR)**. PCR, by then, was a widely used technique in amplifying DNA sequences, and this helped in sequencing to have a clear set of fragments for the final alignment. Another key technique that was employed in mapping known sequences to chromosomes was the **Fluorescence in-situ hybridisation (FISH)** technique.

**FISH Protocol:** In general, FISH used to detect and localize the presence or absence of specific DNA sequences on chromosomes. A probe is first constructed and it is designed to hybridise to its target. The probe is labelled with fluorophores, with targets for antibodies. The labelling can be done in various ways, such as nick translation/PCR using tagged nucleotides. Chromosomes (in metaphase/interphase) are prepared and attached to a glass plate. The DNA probe is then added to the chromosome & incubated for around 12 hours. Antibiotics that specifically identify the fluorophores are added and allowed to hybridise. Using epifluorescence microscopy, the genes are mapped.

[2] Other commonly used vectors are cosmids and yeast artificial chromosomes (YACs).

Overall, the STS data, DNA fingerprint and FISH data accounted for all the 24 different chromosomes observed in homo sapiens (22 autosomes and the X & Y chromosome). The isolated DNA from the above techniques acted as the template for automated Sanger Sequencing and the signals were read off the gels to identify the overall sequence. A series of advancements in DNA sequencing was observed as the HGP was taking shape. They are mentioned in the list below.

1. Instead of four sequencing reactions, one for each primer, there was a switch from dye-labelled primers to dye-labelled terminators. [11]
2. A mutant T7 DNA polymerase was engineered to readily incorporate the above designed dye-labelled terminators. [12]
3. The pre-sequencing steps required the user to ensure that the DNA did not contain contaminants. A magnetic bead-based DNA purification method was designed to automate and thereby simplify the pre-sequencing steps. [13]
4. Capillary electrophoresis eliminated the pouring and loading of gels. It also simplified the extraction and interpretation of the fluorescent signal. [14]
5. Paired-end sequencing helped to link contigs into gapped scaffolds that could be followed up by directed sequencing to close gaps. [15]

The above advancements and the well-established techniques resulted in the successful completion of genome sequencing of *Haemophilus influenzae* [16] (around 2 Mb, 1995) followed quickly by *Saccharomyces cerevisiae* [17] (around 12 Mb, 1996) and *Caenorhabditis elegans* (around 100 Mb, 1998) [18]. By the end of 2004, the entire human genome of a set of people was sequenced as a single mosaic genome that collectively represented all the members of the group.

The **key takeaways** from the HGP are as follows:

1. There are about 22,300 protein-coding genes in human beings.
2. The human genome has a significantly higher number of segmental duplication.
3. Mutation rates are twice as high in males as females.
4. The human genome contains 97% repetitive junk DNA content.
5. Only 2 to 3% of the genome encodes proteins.
6. The human genome has 1.4 million known SNPs, that can be further used to track locations of mutations and so on.

Around the same time, other private research groups also set out to sequence the human genome through other methods. One such group is the **Celera Genomics firm**, led by **Craig Venter**, where they wanted to sequence DNA at a much rapid rate and lower cost. Venter *et al* [19] adapted the technique of **whole-genome shotgun sequencing**, employing **Pairwise end sequencing** [15].

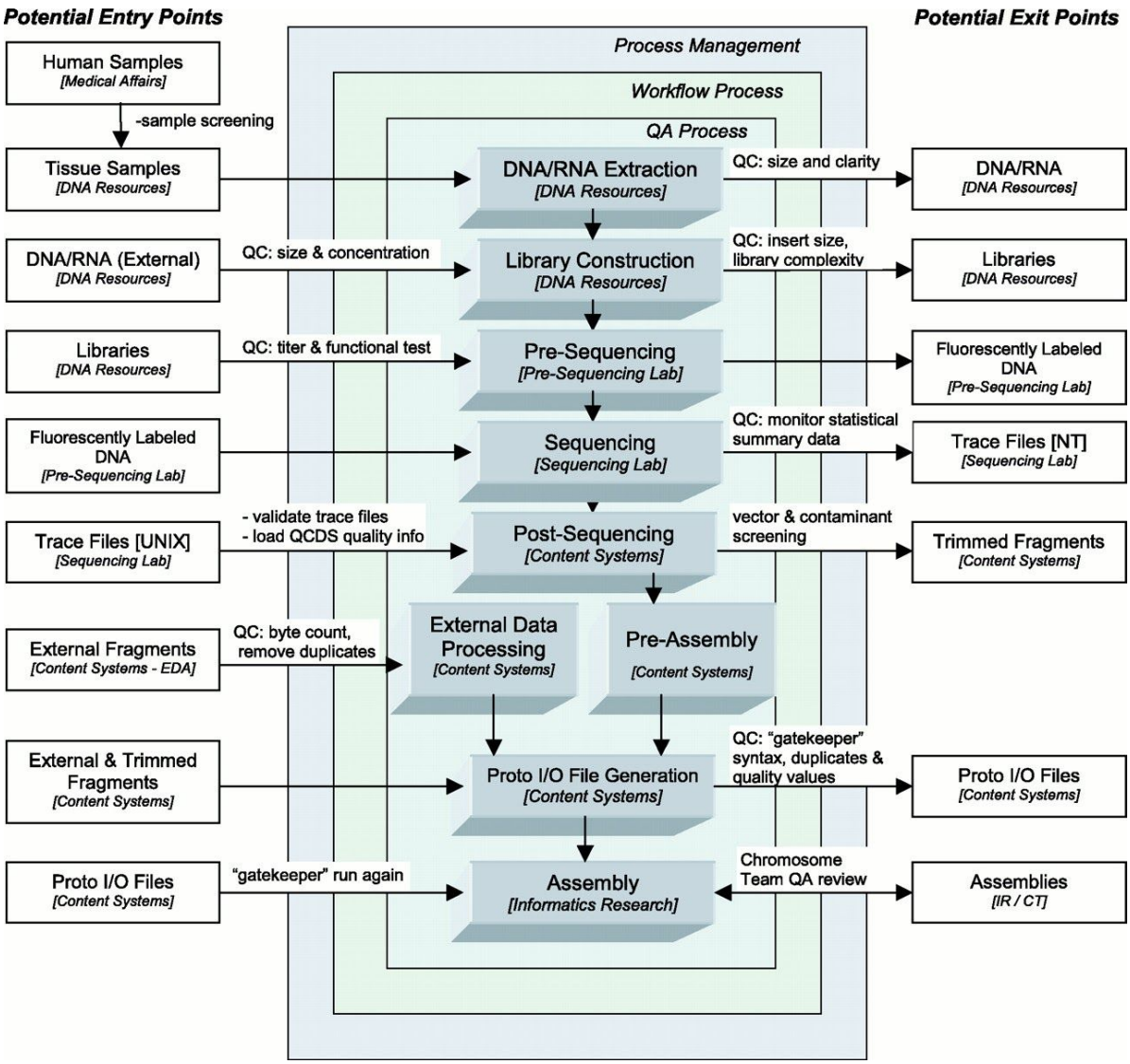


**Figure 3**: Flow diagram explaining all the steps involved in genome sequencing (From Venter *et al*., 2001)

## 3.1 Hierarchical Shotgun Sequencing:

The genome is first broken into larger segments. Once the whole genome is broken down, and the order of the segments is deduced, they are further sheared into small fragments appropriate for sequencing. This method adopts a top-down approach. A low-resolution physical map of the genome is prepared before actual sequencing. From this map, a minimal number of fragments that cover the entire chromosome are selected for sequencing, thereby reducing the quanta of high-throughput sequencing and assembly [20]. The larger fragments are cloned into BAC, and because they were randomly sheared, the ends of each segment stand different and it is theoretically possible to assemble a scaffold of **BAC contigs** that cover the entire genome. This scaffold is called the tiling **path**.

### 3.2 Whole-genome Shotgun Sequencing:

This approach is commonly adopted in the sequencing of genomes of the order 4k Bp to 600k Bp. However, the human genome was of length 3000k Bp. The protocol itself is fairly simple.

1. A high molecular weight DNA sequence is fragmented into smaller parts of length (2k, 10k, 50k, 150k Bp), and cloned into the vector.
2. The clones are then sequenced from both the ends using the chain termination method (Sanger's) to form two short sequences, referred to as *read1* and *read2* individually and as *matepairs* together.
3. The original sequence is then assembled using sequence-assembly software.
    a. The overlapping sequence is collected into longer composite sequences known as *contigs*.
    b. *Contigs* can be linked together as *scaffolds* by identifying the connection between corresponding *matepairs.*
    c. The distance between *contigs* can be inferred from matepairs' positions. Depending on the size between the contigs, different techniques such as PCR (for smaller lengths of 5k-20k Bp) and cloning into BAC (for larger length of >20k Bp).

## 4 Next-generation DNA sequencing - The second generation

The next advent in genome sequencing came about, when the existing methods were optimized to function parallely, therefore quickening the entire process of sequencing and allowing the sequence of longer strands such as the human genome, in one go. This method is referred to as 'Massive-parallel sequencing' or 'Next-generation sequencing' (NGS).

The main difference between the existing electrophoretic techniques and NGS are -

1. Instead of having one tube per each reaction (nucleotide-wise), a complex library of DNA templates is immobilized onto a 2D surface. It is also ensured that the reagent volume can access all the DNA strands.
2. Rather than bacterial cloning, *in vitro* amplification generates copies of each template.
3. Instead of measuring fragment lengths, multiple cycles of biochemistry and imaging ('sequencing-by-synthesis') was introduced.

### 4.1 Sequence-By-Synthesis (SBS)

This technology involves all the DNA strands to be adhered to a flow cell surface in parallel with each other [21]. Four fluorescent-labelled nucleotides are used in this procedure. The labels attached to each nucleotide serve as a terminator for polymerisation. Therefore, after each nucleotide (dNTP) has been incorporated (mediated by DNA polymerase), the fluorescent dye is imaged to find out the

complementary nucleotide to that dNTP thereby identifying the whole sequence. After incorporation, the dye is cleaved off so as to allow the addition of the next nucleotide. The imaging is done at a minimum threshold time-frame so that the whole sequence is covered. This method is highly accurate and shows very little error rate, compared to other technologies.

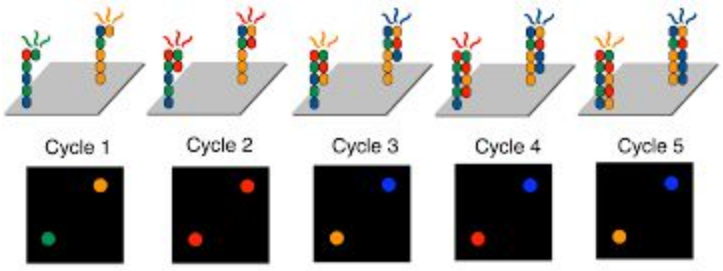


**Figure 4:** Sequence by Synthesis. Incorporation and imaging is shown here. **Source**: http://data-science-sequencing.github.io/Win2018/lectures/lecture2/

## 4.2 Illumina (Solexa) Sequencing

This sequencing technique is based on reversible dye-terminators technology [22], and engineered polymerases. The method works based on 3 simple steps - Amplification, Sequencing and Analyzing.

1. The DNA sequence of our interest is first fragmented to form a genomic library that is easier to handle. This is done by the enzyme - **transposase,** that randomly cleaves the DNA into segments of size 50 - 500 Bp. The other way is to treat the DNA sequence with ultrasonic sound ways (Procedure - **Sonification**). This will also fragment the DNA in similar sizes.
2. An **adapter** sequence is attached to the fragmented DNA either by the transposase itself, or by T7 DNA polymerase and T4 DNA ligase after sonification. Adapters are made of three segments - the sequence complementary to solid support (oligonucleotides on flow cell), the barcode sequence (indices), and the binding site for the sequencing primer. Indices are short sequences of length about 6 bases that act like a tag for each sequence. During analysis, the computer will group the reads of such sequences with the same associated index. The adapters are added on both the sides of the DNA fragments.
3. The sequencing happens on an **acrylamide-coated glass flow cell**. The flow cell has short DNA sequences that serve as a solid support for DNA strands to bind and hold in position. When the fragmented DNA + adapter sequences are washed over the flow cell, the appropriate adapter attaches to the solid support in the cell.
4. The next step is to create larger amounts of the fragmented strand, thus creating a cluster for each unique strand. This is done by **Bridge Amplification**. For a fixed strand, the DNA polymerase moves along the strand to create the complementary strand. The original strand gets washed away in the next step leaving behind the reverse strand. The reverse strand bends and finds a

complementary match to the top adapter sequence. Once that is found, DNA polymerase creates a complementary strand to the reverse strand, thereby creating a copy of the original sequence. The dsDNA is then denatured so that they can act as templates that can get anchored to the flow cell individually.

5. The process of bridge amplification repeats a thousand times to create **Clonal amplification**. Copies of clones are necessary to ensure that smaller errors are averaged and don't cause bigger problems. They are essential for quality control checks. An easy way to check if a sequence is different from the rest in the clone, is by ensuring that the forward and the reverse strands are 100% complementary to each other.
6. Illumina used **SBS technique** for synthesis (explained above).
7. Sequence data is then **analyzed** by finding overlaps in fragments (known as *contigs*) and then lining them accordingly. Repeats in contigs can be very critical to handle, especially in the absence of a reference sequence of the DNA that is being sequenced. Additional difficulties arise with errors from polymerase, PCR-bias, and so on.

Other sequencing approaches in second-generation include that of Pyrosequencing [23], sequencing by reversible terminator chemistry and sequencing-by-ligation mediated by ligase enzymes [24].

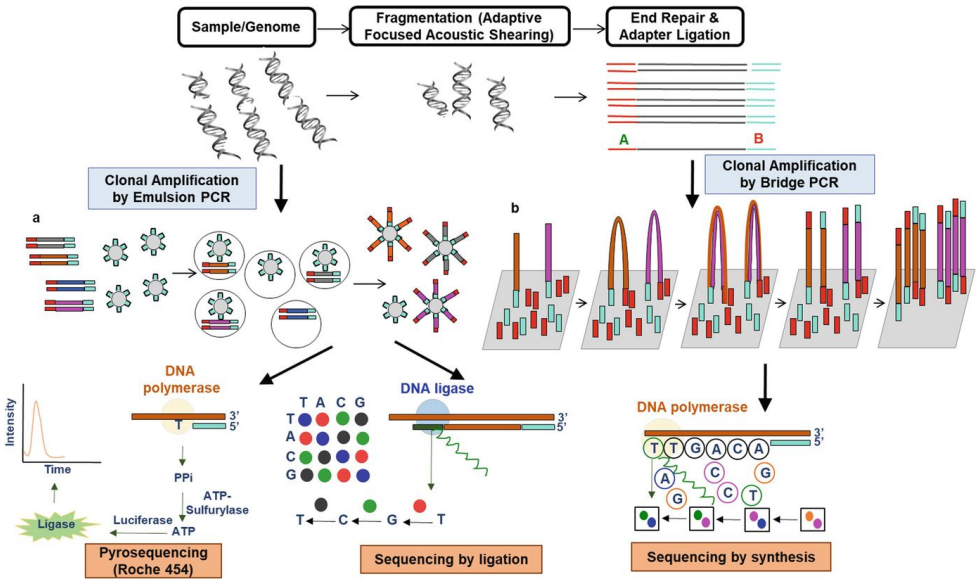


**Figure 5**: NGS Illumina sequencing. **Source**: Nidhi Gupta et al 2019

# 5 Single-Molecule Real-time (SMRT) sequencing

This is a state-of-the-art sequencing technique where DNA strands are read continuously in real-time, rather than having to analyse seq-data after identifying the sequence via one of the above-discussed methods. SMRT utilizes a zero-mode waveguide (ZMW) [25], which is an optical device that directs light into a volume that is smaller in dimension than the wavelength of the light. The PacBio SMRT device consists of multiple ZMWs on a chip. A single DNA polymerase enzyme [26] is fixed at the bottom of ZMW and has a single molecule of DNA strand as the template. Each nucleotide is attached to a unique fluorescent tag for identification purposes. In the SMRT device, when the base is incorporated by the DNA polymerase, the fluorescent tag is cleaved off, and as light passes through the ZMW, the tag diffuses away from the observational area of ZMW. The device is equipped with a detector that detects the fluorescence real-time and identifies the sequence simultaneously. Although this technique is prone to high rates of errors (5-15%), the system can read about 15k to 50 billion Bp from a single device.

Another form of SMRT is the method of **Nanopore sequencing** [27] [28] that is done by recording the change in flow of ions as the nucleotides pass through a nanopore, in the presence of an electric field. A single-stranded DNA fragment is sent through a nanopore and based on the electrostatic interaction of nucleotides with the pore, the sequencer records the plausible nucleotide that was encountered. The machine looks like a regular USB/pen-drive, weighs only 70 grams, and can sequence the entire human genome in less than a day!

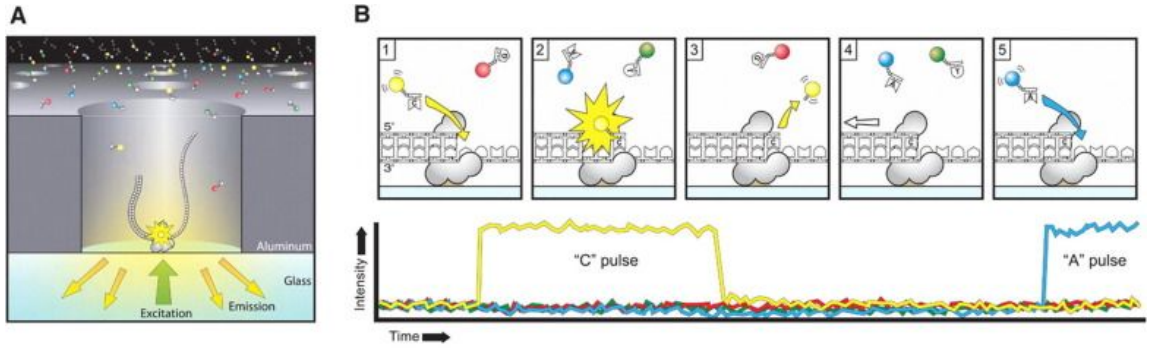


**Figure 6:** Single molecule real-time sequencing. **Source**: PacBio

# 6 Applications of Sequencing

Genome sequencing has been extremely popular and beneficial for understanding the structure and the composition of the genetic material that is the root cause of almost everything. Several **genetic diseases** have been extensively studied with the advent of genome sequencing, with numerous insights offered by the human genome on **genetic variations** that are dependent on various factors such as environment, and background. Of all the various medical applications and diagnostics, **non-invasive prenatal testing** has benefited off the genome sequencing techniques, by the simple ability to count the number of chromosomes to find out chromosomal aneuploidies. Genome sequencing is also leading towards **personalized disease therapy and non-invasive diagnosis**. Shotgun sequencing of complex microbial communities, for example the gut microbiota, will enable deep understanding of oneself and the dynamics of the community as a whole.

# 7 Conclusion

In conclusion, genome sequencing has true potential and can help us uncover a lot of molecular puzzles that were once enigmatic and complex. The various applications of genome sequencing also add to the necessity for future research in this field. Simple techniques such as polymerase chain reaction (PCR), FISH, Chain-termination sequencing and fluorescence dye recognition have played a very major and crucial role in this field.

# 8 References


1. 1958. The Nobel Prize in Chemistry 1958. *NobelPrize.org*
2. **Edman P**, **Högfeldt E**, **Sillén LG**, **Kinell P-O**. 1950. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem. Scand.* **4**: 283–93.
3. **Holley RW, Apgar J, Everett GA, Madison JT,** et al. 1965. Structure of a Ribonucleic Acid. *Science* **147**: 1462–5.
4. **Sanger F**, **Brownlee GG**, **Barrell BG**. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* **13**: 373-IN4.
5. **Wu R**, **Kaiser AD**. 1968. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* **35**: 523–37.
6. **Sanger F**, **Nicklen S**, **Coulson AR**. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**: 5463–7.
7. **admin**Sanger Sequencing: Introduction, Principle, and Protocol | CD Genomics Blog.
8. DNA Sequencing Techniques - Snipcademy.
9. **Staden R**. 1979. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**: 2601–10.
10. **Smith LM, Sanders JZ, Kaiser RJ, Hughes P,** et al. 1986. Fluorescence detection in

automated DNA sequence analysis. *Nature* **321**: 674–9.

11. **Prober JM, Trainor GL, Dam RJ, Hobbs FW,** et al. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**: 336–41.
12. **Tabor S**, **Richardson CC**. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci.* **84**: 4767–71.
13. **DeAngelis MM**, **Wang DG**, **Hawkins TL**. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res.* **23**: 4742–3.
14. **Zhang JianZhong, Fang Yu, Hou JY, Ren HJi,** et al. 1995. Use of Non-Cross-Linked Polyacrylamide for Four-Color DNA Sequencing by Capillary Electrophoresis Separation of Fragments up to 640 Bases in Length in Two Hours. *Anal. Chem.* **67**: 4589–93.
15. **Edwards A, Voss H, Rice P, Civitello A,** et al. 1990. Automated DNA sequencing of the human HPRT locus. *Genomics* **6**: 593–608.
16. **Fleischmann RD, Adams MD, White O, Clayton RA,** et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496–512.
17. **Goffeau A, Barrell BG, Bussey H, Davis RW,** et al. 1996. Life with 6000 Genes. *Science* **274**: 546–67.
18. **C. elegans Sequencing Consortium**. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**: 2012–8.
19. **Venter JC, Adams MD, Myers EW, Li PW,** et al. 2001. The Sequence of the Human Genome. *Science* **291**: 1304–51.
20. **Ronsisvalle L**A Primer of Genome Science EDITION.
21. **Balasubramanian S**, **Klenerman D**, **Barnes C**. 2003. Arrayed polynucleotides and their use in genome analysis.
22. **Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP,** et al. 2008. Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* **456**: 53–9.
23. **Pettersson E**, **Lundeberg J**, **Ahmadian A**. 2009. Generations of sequencing technologies. *Genomics* **93**: 105–11.
24. **Huang Y-F, Chen S-C, Chiang Y-S, Chen T-H,** et al. 2012. Palindromic sequence impedes sequencing-by-ligation mechanism. *BMC Syst. Biol.* **6**: S10.
25. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations | Science.
26. **Eid J, Fehr A, Gray J, Luong K,** et al. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**: 133–8.
27. **Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H,** et al. 2014. Decoding long nanopore sequencing reads of natural DNA. *Nat. Biotechnol.* **32**: 829–33.
28. **Branton D, Deamer DW, Marziali A, Bayley H,** et al. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**: 1146–53.