Swayam Course - Analytical Techniques

Week 13, Tutorial 35 - Introduction to Genomic techniques

Content Writer - Dr. Ruby Dhar, DBT-Biocare Fellow, Department of Biochemistry, All India Institute of Medical Sciences, New Delhi

Introductions

An organism's complete set of DNA is called its genome. Virtually every single cell in the body contains a complete copy of the approximately 3 billion DNA base pairs, or letters, that make up the human genome.

Deoxyribonucleic acid (DNA) is the chemical compound that contains the instructions needed to develop and direct the activities of nearly all living organisms. DNA molecules are made of two twisting, paired strands, often referred to as a double helix

Each DNA strand is made of four chemical units, called nucleotide bases, which comprise the genetic "alphabet." The bases are adenine (A), thymine (T), guanine (G), and cytosine (C). Bases on opposite strands pair specifically: an A always pairs with a T; a C always pairs with a G. The order of the As, Ts, Cs and Gs determines the meaning of the information encoded in that part of the DNA molecule just as the order of letters determines the meaning of a word. With its four-letter language, DNA contains the information needed to build the entire human body. A gene traditionally refers to the unit of DNA that carries the instructions for making a specific protein or set of proteins. Each of the estimated 20,000 to 25,000 genes in the human genome codes for an average of three proteins.

Located on 23 pairs of chromosomes packed into the nucleus of a human cell, genes direct the production of proteins with the assistance of enzymes and messenger molecules. Specifically, an enzyme copies the information in a gene's DNA into a molecule called messenger ribonucleic acid (mRNA). The mRNA travels out of the nucleus and into the cell's cytoplasm, where the mRNA is read by a tiny molecular machine called a ribosome, and the information is used to link together small molecules called amino acids in the right order to form a specific protein.

The understanding of the DNA sequence is done by using various tools of genomics. This module will introduce to the various genomic techniques.

OBJECTIVES:

1. **WHAT IS A GENOME**
2. **WHAT IS GENOMICS**
3. **BASIC METHOGOLOGY TO STUDY GENOMICS**
4. **INTRODUCTION TO DNA SEQUENCING**
5. **SERIAL ANALYSIS OF GENE EXPRESSION**
6. **MICROARRAYS**
7. **NEXT GENERATION SEQUENCING**
8. **CHROMATIN IMMUNOPRECIPITATION**
9. **GENOME EDITING**
10. **SUMMARY**

Before we discuss the techniques to understand the genomics, I would like to introduce to the common terms used in genomics. So we start with the understanding of the genome.

A genome is an organism's complete set of genetic instructions. Each genome contains all of the information needed to build that organism and allow it to grow and develop.

What is a genome?

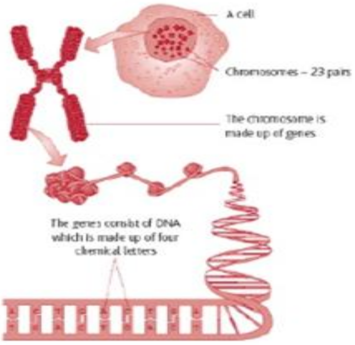
The genome broadly refers to the total amount of DNA of a single cell.

(Haploid cell in the case of a diploid organism)

DNA encodes the whole hereditary information of an organism.

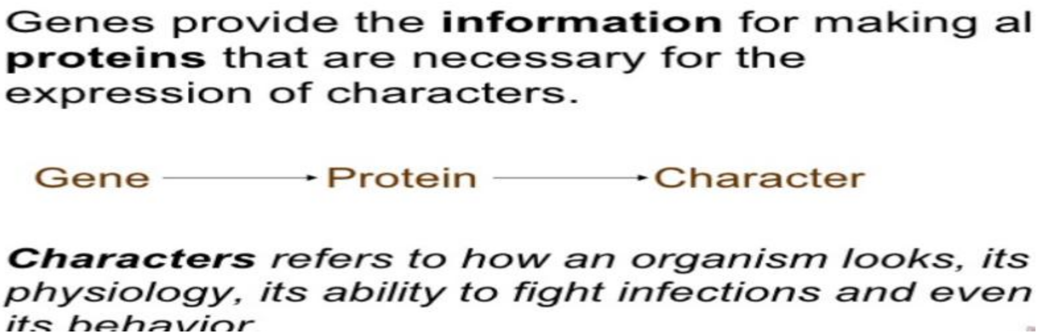
The instructions in our genome are made up of DNA.

Representation of a human cell and nucleus with DNA


A cell
Chromosomes – 23 pairs
The chromosome is made up of genes
The genes consist of DNA which is made up of four chemical letters

The human genome is made of 3.2 billion bases of DNA but other organisms have different genome sizes.

- Within DNA is a unique chemical code that guides our growth, development and health.
- This code is determined by the order of the four nucleotide bases that make up DNA, adenine, cytosine, guanine and thymine, A, C, G and T for short.



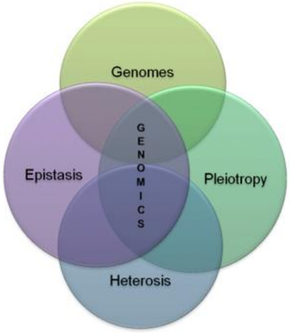
Genes provide the **information** for making al **proteins** that are necessary for the expression of characters.

Gene → Protein → Character

Characters *refers to how an organism looks, its physiology, its ability to fight infections and even its behavior*

- Genomics is the study of whole genomes of organisms and incorporates elements from genetics.
- Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes.
- It differs from 'classical genetics' in that it considers an organism’s full complement of hereditary material, rather than one gene or one gene product at a time.
- Genomics also focuses on interactions between loci and *alleles* within the genome and other interactions such as *epistasis, pleiotropy* and *heterosis*.
- Genomics harnesses the availability of complete DNA sequences for entire organisms and was made possible by both the pioneering work of Fred Sanger and the more recent next-generation sequencing technology.

What is Genomics?



Epistasis= the interaction between genes

Heterosis= hybrid vigour

Pleiotropy= where one gene affects multiple characteristics.

BASIC METHODOLOGY TO STUDY THE GENOME

(1) Genomic DNA Isolation

(2) Separation of DNA

(3) Cutting and Joining of DNA

(4) Cloning and Vectors

(5) Detection of Gene of Interest

(6) Recombinant DNA and Cloning

(7) Production of Multiple Copies of DNA, Using Polymerase Chain Reaction (PCR)

(8) DNA Sequencing-Sanger /NGS

(9) Transcriptome analysis: Microarray/NGS

(10) Chromatin Immunoprecipitation followed by PCR, microarray/NGS

(11) Genome editing

INTRODUCTION TO DNA SEQUENCING

The Techniques related to DNA isolation and recombinant DNA technology has been discussed in another module, hence this module will focus on other methods, from DNA sequencing techniques.

Early efforts at sequencing genes were painstaking, time consuming, and labor intensive, such as when Gilbert and Maxam reported the sequence of 24 base pairs using a method known as wandering-spot analysis (Gilbert & Maxam, 1973). Thankfully, this situation began to change during the mid-1970s, when researcher Frederick Sanger developed several faster, more efficient techniques to sequence DNA. Indeed, Sanger's work in this area was so groundbreaking that it led to his receipt of the Nobel Prize in Chemistry in 1980. Over the next several decades, technical advances automated, dramatically sped up, and further refined the Sanger sequencing process. Also called the chain-termination or dideoxy method, Sanger sequencing involves using a purified DNA polymerase enzyme to synthesize DNA chains of varying lengths.

In 1986, a company named Applied Biosystems began to manufacture automated DNA sequencing machines based on the Sanger method. These machines used fluorescent dyes to tag each nucleotide, allowing the reactions to be run in one column and read by color (rather than by what lane they appeared in). By running 24 samples at a time, the $100,000 machines yielded 12,000 "letters" of DNA per day. Over time, sequencing technology (and DNA synthesis technology) advanced with more sophisticated separation strategies, alternative visualization strategies, and more parallel samples. As a result, today's machines can typically handle 96 samples at a time. In addition, whereas traditional, gel-based Sanger sequencing and early machines could only generate 250 to 500 base pairs of DNA sequence per reaction, 750 to 1,000 base pairs of sequence can now be read from a single reaction, making sequencing a much less expensive option than it used to be

Less expensive.

Technical Breakthrough For DNA Sequencing

In 1977, two separate methods for the large-scale sequencing of DNA were devised:

- ❑ **Chemical cleavage method**

  by A. M. Maxam and W. Gilbert

- ❑ **Enzymatic chain termination method**

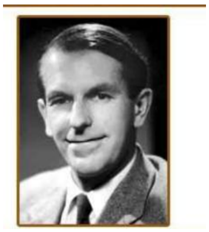
  By F. Sanger et al

Allan Maxam

Walter Gilbert

Recipient of two Noble prize:

- **1958 for his discovery of the structure of the insulin molecule**
- **1980 for his collaborative work on base sequences in nucleic acids with Paul Berg and Walter Gilbert.**



F
rederick Sanger

Maxam-Gilbert method for sequencing (Chemical Cleavage method)

This method is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides.

The purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are hydrolysed using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the reaction of thymine for the C-only reaction

Sequential steps involved in the process

label 5’end of DNA

Aliqot DNA sanple in Tube

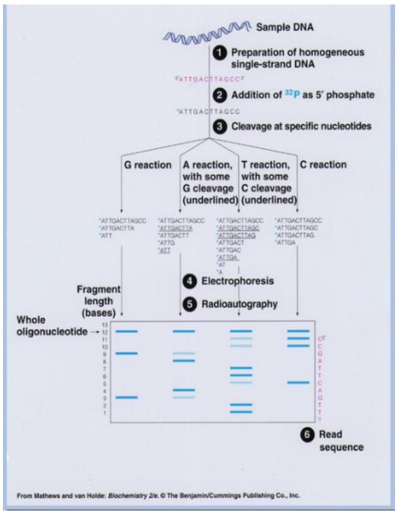
Perform base modification reaction

Perform cleavage reaction

Perform Gel Electropheresis

Perform Auroradiography

Interpret results



From Mathews and van Holde: *Biochemistry 2/e*. © The Benjamin/Cummings Publishing Co., Inc.

Sanger Sequencing (Enzymatic chain termination method)

This method uses single-stranded DNA.

Also known as **dideoxy sequencing** method because it involves the use of analogue of normal nucleotide 2’,3’-dideoxynucleoside triphosphates (ddNTPs).

The 3’-OH group necessary for formation of the phosphodiester bond is missing in ddNTPs

This method is based upon the incorporation of ddNTPs into a growing DNA strand to stop chain elongation.

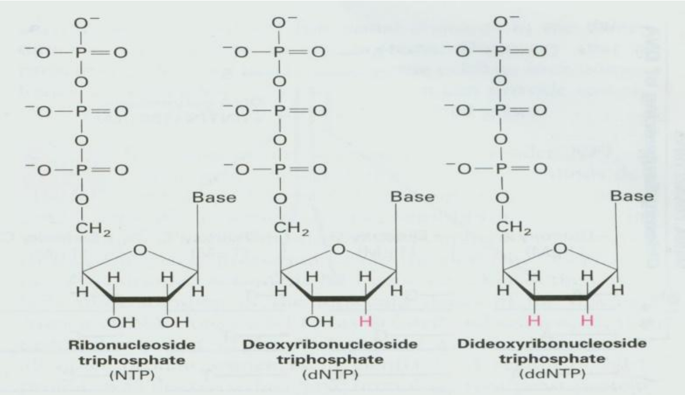
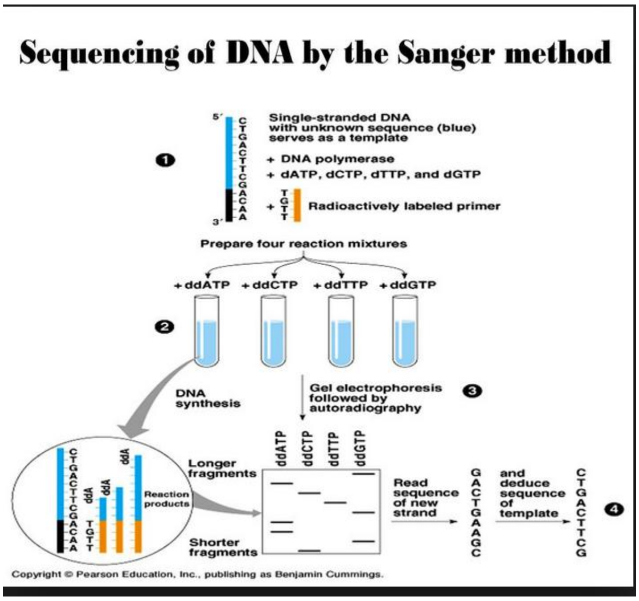

Base
CH2
OH OH
H
Ribonucleoside triphosphate (NTP)
Deoxyribonucleoside triphosphate (dNTP)
Dideoxyribonucleoside triphosphate (ddNTP)

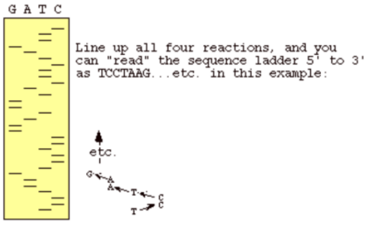
Figure: Structure of NTP, dNTP, and ddNTP

Lodish, H.;Berk, A. *et. al.* (4th ed);

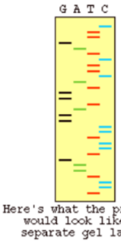
Mol. Cell Biol.; W. H. Freeman and Co. (2000), p: 233

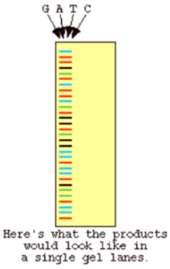
Figure showing the basic steps involved in Sanger Sequencing


Sequencing of DNA by the Sanger method
5'
3'
Single-stranded DNA with unknown sequence (blue) serves as a template
+ DNA polymerase
+ dATP, dCTP, dTTP, and dGTP
Radioactively labeled primer
1
Prepare four reaction mixtures
+ddATP +ddCTP +ddTTP +ddGTP
2
DNA synthesis
Gel electrophoresis followed by autoradiography
3
Longer fragments
Reaction products
Shorter fragments
ddATP ddCTP ddTTP ddGTP
Read sequence of new strand
and deduce sequence of template
4
Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.


G A T C
Line up all four reactions, and you can "read" the sequence ladder 5' to 3' as TCCTAAG...etc. in this example:
etc.

In automated sequencing, the oligonucleotide primers can be "end-labeled" with different color dyes, one foreach ddNTP.


G A T C
Here's what the products would look like in separate gel lanes.


G A T C
Here's what the products would look like in a single gel lanes.

These dyes fluoresce at different wave-lengths

Difference between Sanger and Maxam Gilbert Method:

SANGER METHOD	**MAXAM GILBERT METHOD**
Enzymatic	Chemical
Requires DNA synthesis	Requires DNA
Termination of chain elongation	Breaks DNA at different

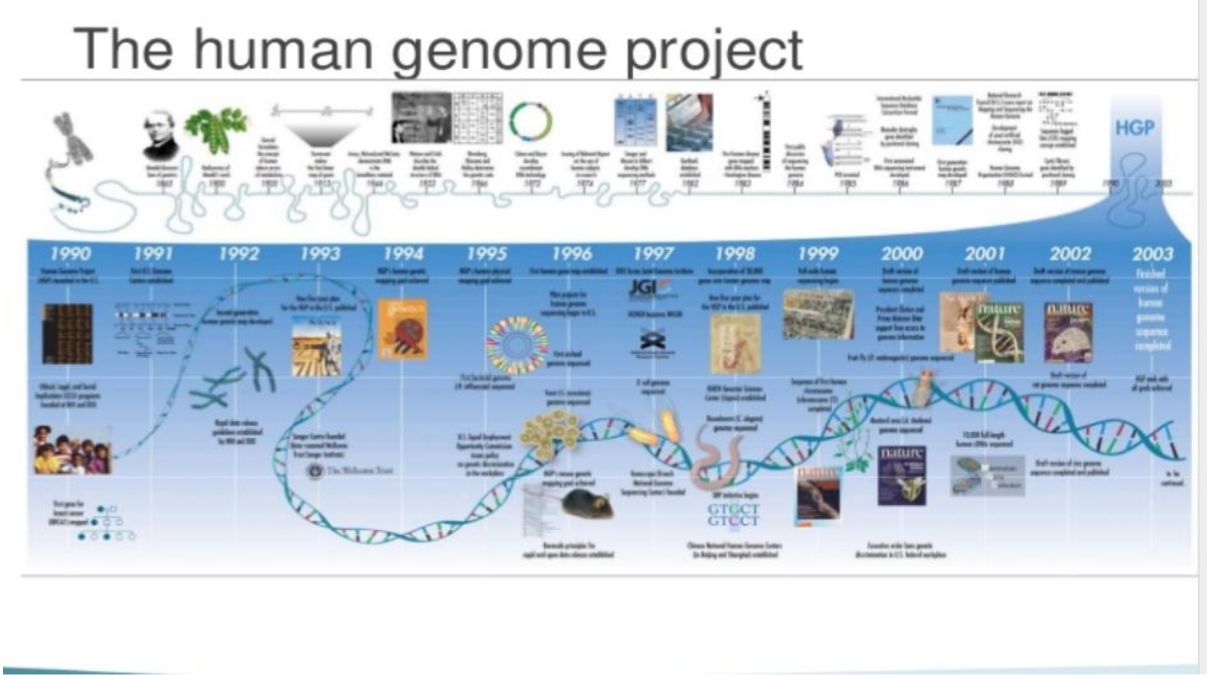
	Nucleotides
Automation	Automation is not available
Single Stranded DNA	Single stranded Stranded DNA

Maxam –Gilbert method was published two years after the Sanger method, but rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each DNA to be cloned for production of single-stranded DNA.

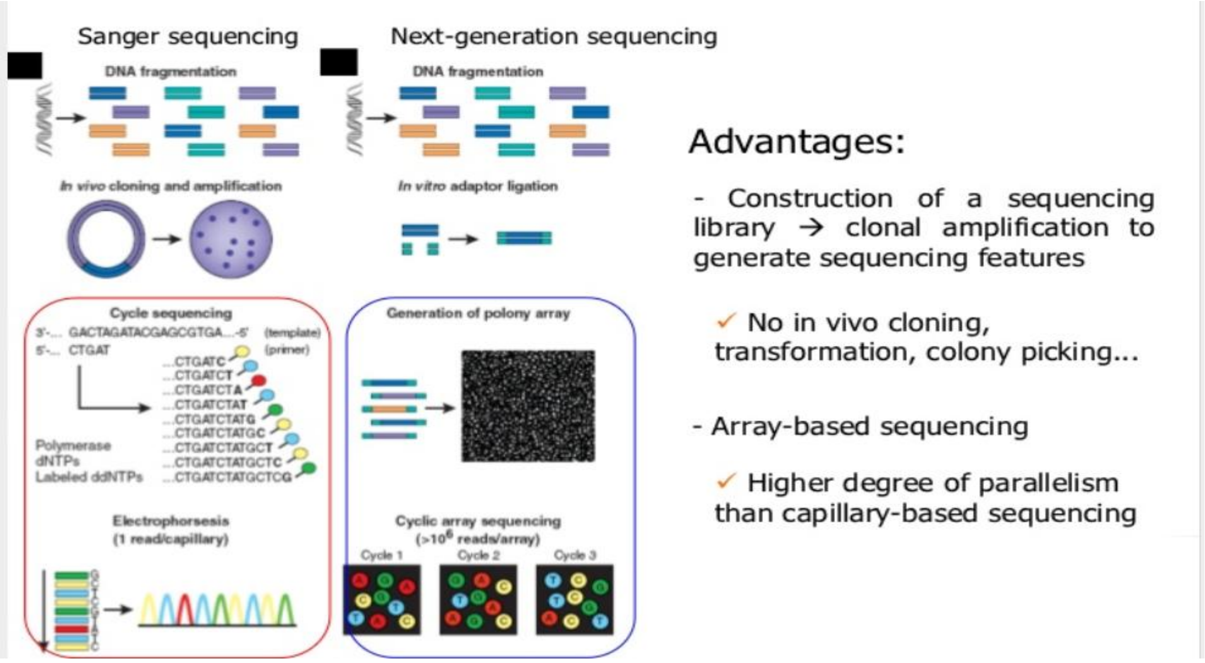
The Human genome project

The main goals of the Human Genome Project were first articulated in 1988 by a special committee of the U.S. National Academy of Sciences, and later adopted through a detailed series of five-year plans jointly written by the National Institutes of Health and the Department of Energy. James Watson was appointed to lead the NIH component, which was dubbed the Office of Human Genome Research. The following year, the Office of Human Genome Research evolved into the National Center for Human Genome Research.

In 1990, the initial planning stage was completed with the publication of a joint research plan, "Understanding Our Genetic Inheritance: The Human Genome Project, The First Five Years, FY 1991-1995." HGP researchers deciphered the human genome in three major ways: determining the order, or "sequence," of all the bases in our genome's DNA; making maps that show the locations of genes for major sections of all our chromosomes; and producing what are called linkage maps, through which inherited traits (such as those for genetic disease) can be tracked over generations.



Sanger Chain termination 1977



Advantages:

- Construction of a sequencing library → clonal amplification to generate sequencing features

✓ No in vivo cloning, transformation, colony picking...

- Array-based sequencing

✓ Higher degree of parallelism than capillary-based sequencing

Next-generation sequencing instruments can generate as much data in one day as several hundred Sanger-type

What is next generation sequencing??

With the beginning of the automated Sanger method, the first generation of high throughput sequencing began. But then the sequencing technology evolved to sequence genomes at a much faster rate and in a more efficient manner. Next generation sequencing (NGS), massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionised genomic research. Using NGS an entire human genome can be sequenced within a single day. In contrast, the previous Sanger sequencing technology, used to decipher the human genome, required over a decade to deliver the final draft.

There are a number of different NGS platforms using different sequencing technologies, which are discussed in this module. However, all NGS platforms perform sequencing of millions of small fragments of DNA in parallel. Bioinformatics analyses are used to piece together these fragments by mapping the individual reads to the human reference genome. Each of the three billion bases in the human genome is sequenced multiple times, providing high depth to deliver accurate data and an insight into unexpected DNA variation.

Various genomic techniques will be discussed in this module:

Functional genomics is the study of gene function through parallel expression measurements of a genome.

DNA microarrays

oligonucleotide microarrays

Serial analysis of gene expression (SAGE)

Next Generation

sequencing

Genome editing

SAGE: SERIAL ANALYSIS OF GENE EXPRESSION

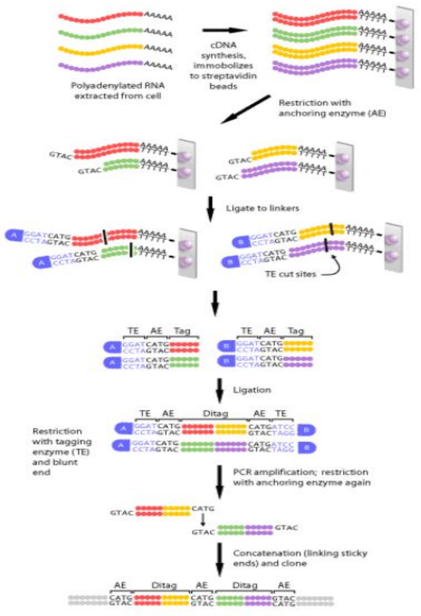
SAGE is invented by Dr Victor Velculescu in 1995 at John Hopkins University, USA

SAGE allows   rapid and detailed analysis of overall gene expression patterns

The basic concept of SAGE rests on two principles:

- A small sequence of nucleotides from the transcript, called a 'tag', can effectively identify the original transcript from whence it came
- linking these tags allows for rapid sequencing analysis of multiple transcripts.

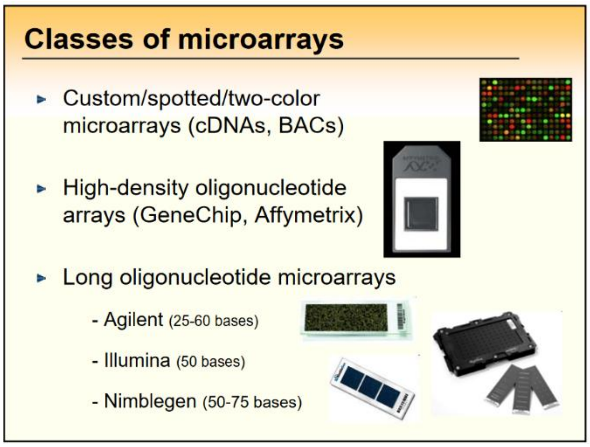
One major advantage of SAGE is it doesn't require prior knowledge of sequence of DNA/RNA


AAAAA
Polyadenylated RNA extracted from cell
cDNA synthesis, immobolizes to streptavidin beads
Restriction with anchoring enzyme (AE)
GTAC
Ligate to linkers
A
B
GGATCATG
CCTAGTAC
TE cut sites
TE
AE
Tag
Ligation
Ditag
Restriction with tagging enzyme (TE) and blunt end
CATGATCC
GTACTAGG
PCR amplification; restriction with anchoring enzyme again
CATG
GTAC
Concatenation (linking sticky ends) and clone
AE
Ditag

MICROARRAYS:

A DNA *microarray* (also commonly known as DNA chip or biochip) is a collection of microscopic DNA spots attached to a solid surface.

Specific sequences are immobilized to a surface and reacted with labeled cDNA targets. A signal resulting from hybridization of the labeled target with the specific immobilized probe identifies which RNAs are present in the unknown target sample.



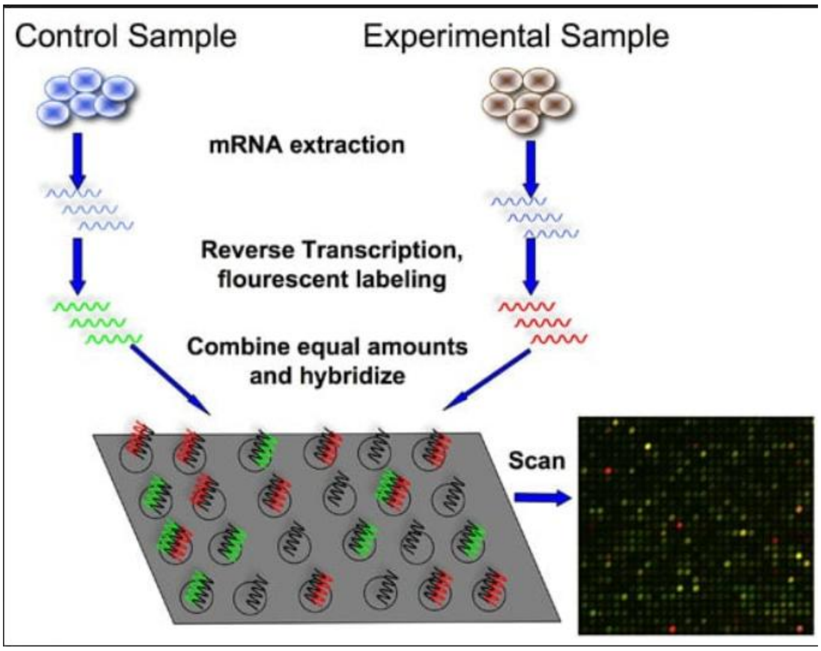
Classes of microarrays

- Custom/spotted/two-color microarrays (cDNAs, BACs)
- High-density oligonucleotide arrays (GeneChip, Affymetrix)
- Long oligonucleotide microarrays
  - Agilent (25-60 bases)
  - Illumina (50 bases)
  - Nimblegen (50-75 bases)

Spotting of DNAs at high density onto a glass microscopy slide (5000-10000 spots per slide) and cross-linking to the glass surface

Two independent mRNA or DNA samples are fluorescently labeled with Cy3 (green) or Cy5 (red

A laser scans the slide and calculates the ratio of fluorescence intensities between the two samples.


Control Sample
Experimental Sample
mRNA extraction
Reverse Transcription, flourescent labeling
Combine equal amounts and hybridize
Scan

One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. DNA Microarray technology has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body.

Applications of Microarrays

Gene Discovery: DNA Microarray technology helps in the identification of new genes, know about their functioning and expression levels under different conditions.

Disease Diagnosis: DNA Microarray technology helps researchers learn more about different diseases such as heart diseases, mental illness, infectious disease and especially the study of cancer. Until recently, different types of cancer have been classified on the basis of the organs in which the tumors develop. Now, with the evolution of microarray technology, it will be possible for the researchers to further classify the types of cancer on the basis of the patterns of gene activity in the tumor cells. This will tremendously help the pharmaceutical community to develop more effective drugs as the treatment strategies will be targeted directly to the specific type of cancer.

Drug Discovery: Microarray technology has extensive application in *Pharmacogenomics.* Pharmacogenomics is the study of correlations between therapeutic responses to drugs and the genetic profiles of the patients. Comparative analysis of the genes from a diseased and a normal cell will help the identification of the biochemical constitution of the proteins synthesized by the diseased genes.

Toxicological Research: Microarray technology provides a robust platform for the research of the impact of toxins on the cells and their passing on to the progeny. Toxicogenomics establishes correlation between responses to toxicants and the changes in the genetic profiles of the cells exposed to such toxicants.

NEXT GENERATION SEQUENCING (SECOND GENERATION)

Deoxyribonucleic acid, commonly known as DNA, contains the blueprints of life. Within its structures are the codes required for the assembly of proteins and non-coding RNA – these molecular machineries affect all the biological systems that create and maintain life. By understanding the sequence of DNA, researchers have been able to elucidate the structure and function of proteins as well as RNA and have gained an understanding of the underlying causes of disease. Next Generation Sequencing (NGS) is a powerful platform that has enabled the sequencing of thousands to millions of DNA molecules simultaneously. This powerful tool is revolutionizing fields such as personalized medicine, genetic diseases, and clinical diagnostics by offering a high throughput option with the capability to sequence multiple individuals at the same time.

Here is a list of commonly used words used in sequencing based on NGS platforms:

Reads:

The output of an NGS sequencing reaction. A read is a single uninterrupted series of nucleotides representing the sequence of the template.

Read Length: The length of each sequencing read. This variable is always represented as an average read length since individual reads have varying lengths.

Coverage:

The number of times a particular nucleotide is sequenced. Due to the error -prone sequencing reactions, random errors could occur. Therefore, 30x coverage is typically required to ensure each nucleotide sequence is accurate.

Deep Sequencing:

Sequencing where the coverage is greater than 30x. This is used in cases where dealing with rare polymorphisms which only a subset of the sample expresses the mutation. This method increases range, complexity, sensitivity, and accuracy of the result.

Paired-End Sequencing:

Sequencing from both ends of a fragment while keeping track of the paired data. With this method the sequencing reaction will commence from one end of the fragment. Once completed, the fragment is denatured and a sequencing primer is hybridized to the reverse side adapter. The fragment is then sequenced again. Using this method will allow either further confirmation of the accuracy of the sequence or it could be used to increase the overall read length.

Mate-Paired reads:

A sample preparation step where large DNA fragments (~10kb) are circularized with an adapter sequence followed by degradation of the circular DNA. This method links DNA fragments that are separated from each other by a certain distance and it is used in applications such as *de novo* assembly, structural variant detection, and identification of complex genomic rearrangements.

Adapter:

Unique sequences used to cap the ends of a fragmented DNA. The adapter's functions are as follows: 1) allow hybridization to solid surface; 2) provide priming location for both amplification and sequencing primers; and 3) provide barcoding for multiplexing different samples in the same run.

Library:

A collection of DNA fragments with adapters ligated to each end. Library preparation is required before a sequencing run. Our next knowledge base will delve into the different sample and library preparation methods available.

Alignment:

Mapping a sequence read to a known reference genome

Reference sequence/genome:

A fully sequenced and mapped genome used for the mapping of sequence reads.

De Novo Assembly:

Assembly of the sequence reads to generate a reference sequence.

Specificity:

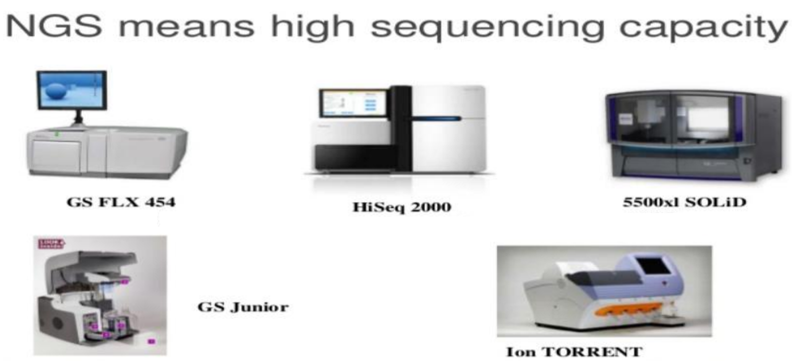
The percentage of sequences that map to the intended targets out of total bases per run.

Uniformity:

The variability in sequence coverage across target regions. When performing whole genome sequencing or exome sequencing, it is expected that the result will be highly uniform (as there should be a 1:1 ratio in the starting material). However, RNA sequencing will not be uniform since differences in expression alter its starting material.

Homopolymer:

A stretch of single nucleotide bases, such as AAAA or GGGGGG.

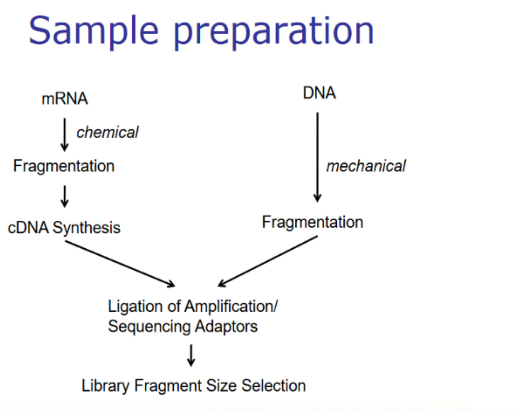


NGS means high sequencing capacity

All NGS based sequencing are based on the following features:

1. Massively parallel sequencing :
   Sequencing by synthesis
   Sequencing by ligation
2. Mostly produce short reads from <400bp
3. Read numbers vary from 1million to 1Billion per run

Before we discuss various sequencing platforms, here are steps involved before loading the sample to the sequencer



Sample preparation

mRNA → chemical → Fragmentation → cDNA Synthesis

DNA → mechanical → Fragmentation

Ligation of Amplification/ Sequencing Adaptors

Library Fragment Size Selection

454 Sequencing

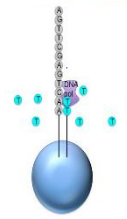
Sequencing by 454 is based on pyrosequencing .

- Read length 250-500 bp .
- Generates more than >1 million reads per run
- Roche 454 sequencing can sequence much longer reads than Illumina.
- Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added.

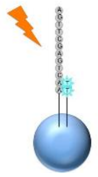
- DNA or RNA is fragmented into shorter reads, in this case up to 1kb.
- Generic adaptors are added to the ends and these are annealed to beads, one DNA fragment per bead. The fragments are then amplified by PCR

Each bead is then placed in a single well of a slide. So each well will contain a single bead, covered in many PCR copies of a single sequence. The wells also contain DNA polymerase and sequencing buffers.

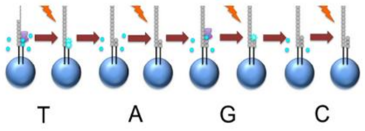
The slide is flooded with one of the four NTP species. Where this nucleotide is next in the sequence, it is added to the sequence read.

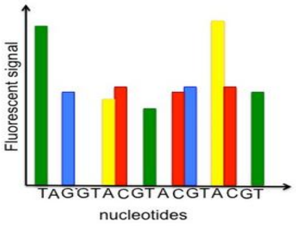


The addition of each nucleotide releases a light signal. These locations of signals are detected and used to determine which beads the nucleotides are added to.



This NTP mix is washed away. The next NTP mix is now added and the process repeated, cycling through the four NTPs.





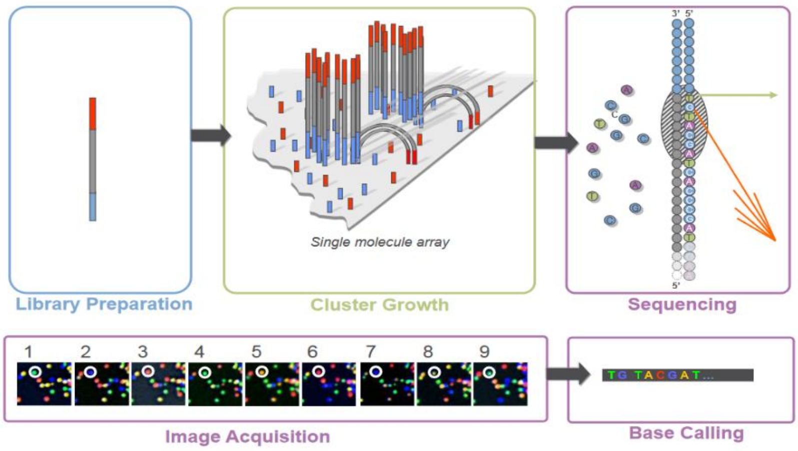
The final sequence from the instrument

SOLEXA GENOME ANALYZER

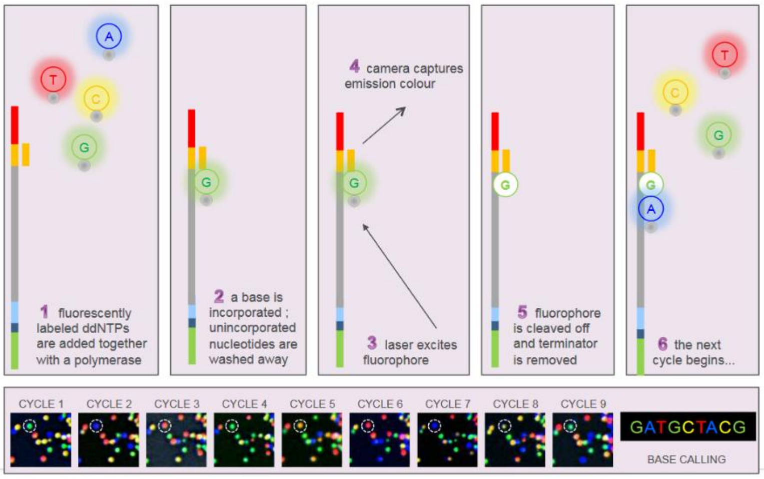
- Chemistry based on reversible terminators
- Sample amplified by solid-phase amplification
- Read length 30-75 bp
- >100 million reads per run
- ~10 Gb of sequence

4-8 days run

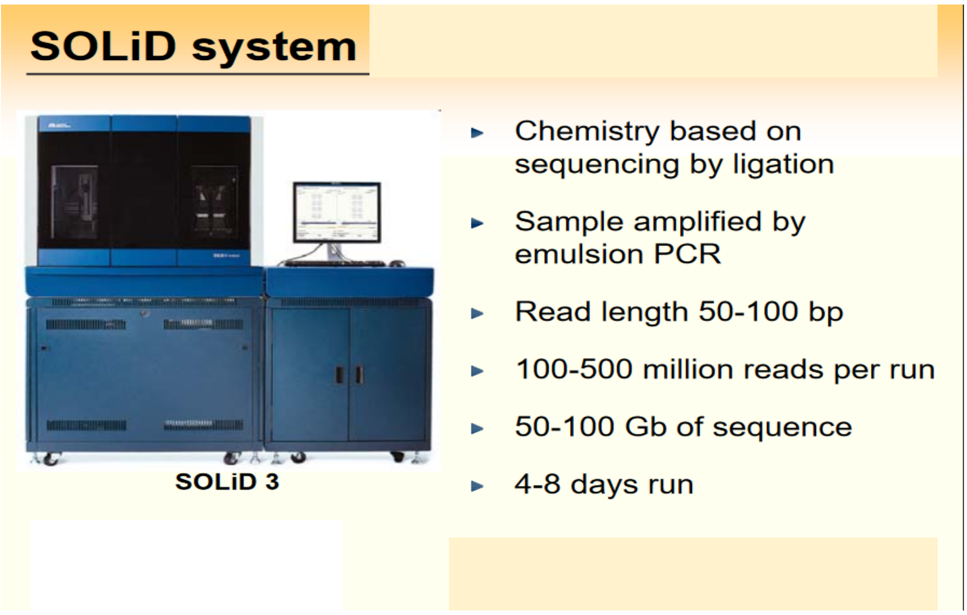
Illumina Sequencing Workflow



Sequencing by synthesis (SBS) close-up


A
T
C
G
1 fluorescently labeled ddNTPs are added together with a polymerase
2 a base is incorporated ; unincorporated nucleotides are washed away
4 camera captures emission colour
3 laser excites fluorophore
5 fluorophore is cleaved off and terminator is removed
6 the next cycle begins...
CYCLE 1
CYCLE 2
CYCLE 3
CYCLE 4
CYCLE 5
CYCLE 6
CYCLE 7
CYCLE 8
CYCLE 9
GATGCTACG
BASE CALLING

Sequencing by synthesis utilizes the step-by-step incorporation of reversibly fluorescent and terminated nucleotides for DNA sequencing and is used by the Illumina NGS platforms. The nucleotides used in this method have been modified in two ways: 1) each nucleotide is reversibly attached to a single fluorescent molecule with unique emission wavelengths, and 2) each nucleotide is also reversibly terminated ensuring that only a single nucleotide will be incorporated per cycle. All four nucleotides are added to the sequencing chip and after nucleotide incorporation the remaining DNA bases are washed away. The fluorescent signal is read at each cluster and recorded; both the fluorescent molecule and the terminator group are then cleaved and washed away. This process is repeated until the sequencing reaction is complete. This system is able to overcome the disadvantages of the pyrosequencing system by only incorporating a single nucleotide at a time

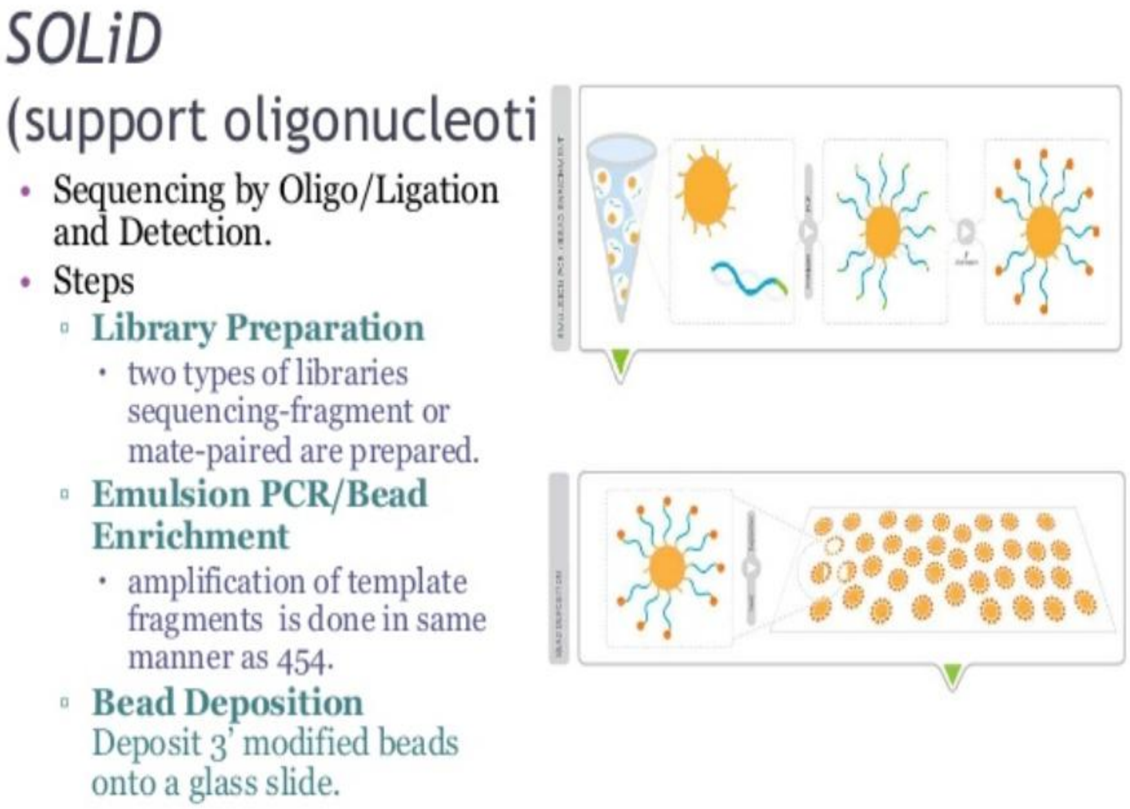


SOLiD system

SOLiD 3

- Chemistry based on sequencing by ligation
- Sample amplified by emulsion PCR
- Read length 50-100 bp
- 100-500 million reads per run
- 50-100 Gb of sequence
- 4-8 days run

Sequencing by ligation is different from the other two methods since it does not utilize a DNA polymerase to incorporate nucleotides. Instead, it relies on short oligonucleotide probes that are ligated to one another. These oligonucleotides consist of 8 bases (from 3’-5’): two probe specific bases (there are a total of 16 8-mer probes which all differ at these two base positions) and six degenerate bases; one of four fluorescent dyes are attached at the 5’ end of the probe. The sequencing reaction commences by binding of the primer to the adapter sequence and then hybridization of the appropriate probe. This hybridization of the probe is guided by the two probe specific bases and upon annealing, is ligated to the primer sequence through a DNA ligase. Unbound oligonucleotides are washed away, the signal is detected and recorded, the fluorescent signal is cleaved (the last 3 bases), and then the next cycle commences. After approximately 7 cycles of ligation the DNA strand is denatured and another sequencing primer, offset by one base from the previous primer, is used to repeat these steps - in total 5 sequencing primers are used.



SOLiD

(support oligonucleoti

- Sequencing by Oligo/Ligation and Detection.
- Steps
  - Library Preparation
    - two types of libraries sequencing-fragment or mate-paired are prepared.
  - Emulsion PCR/Bead Enrichment
    - amplification of template fragments is done in same manner as 454.
  - Bead Deposition Deposit 3' modified beads onto a glass slide.

Third-generation sequencing: Emerging technologies for single-molecule sequencing

Third-generation single-molecule sequencing technologies have emerged to reduce the price of sequencing and to simplify the preparatory procedures and sequencing methods

Third generation sequencing works by reading the nucleotide sequences at the single molecule level, in contrast to existing methods that require breaking long strands of DNA into small segments then inferring nucleotide sequences by amplification and synthesis.

Third generation sequencing technologies have the capability to produce substantially longer reads than second generation sequencing.

However third generation sequencing has error rates at almost unrepairable levels, rendering the technologies impractical for certain applications such as de novo genome assembly

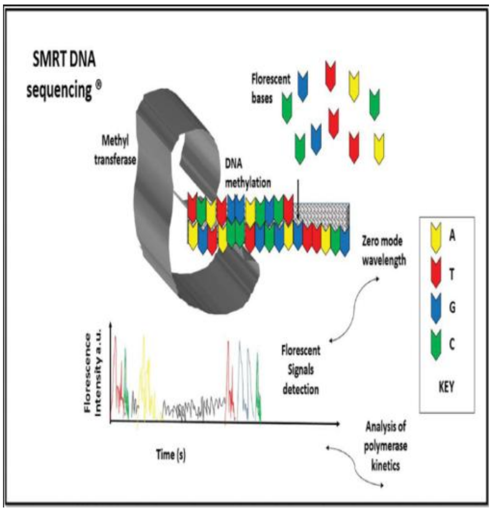
Three main advantage of NGS:

a.Longer reads

b.Epigenetics

c.Portability and speed

Single molecule real time sequencing (SMRT)



One molecule of DNA polymerase is immobilized at the bottom of each well using the biotin-streptavidin system in nanostructures known as zero-mode waveguides (ZMWs).

Once the template single-strand DNA is coupled with immobilized DNA polymerase, fluorescently labeled dNTP analogs are added and detected when the nucleotide is incorporated into the growing strand.

CCD cameras continuously monitor the 150,000 ZMWs as a series of observed pulses that are converted into single molecular traces representing the template sequences.

All four nucleotides are added simultaneously and measured in real time, the speed of sequencing is much increased.

Third-generation sequencing:

- **Nanopore sequencing (MinION and PromethION)**

The MinION Mkl is a portable handheld device for DNA and RNA sequencing that attaches directly to a laptop/computer using a USB port.

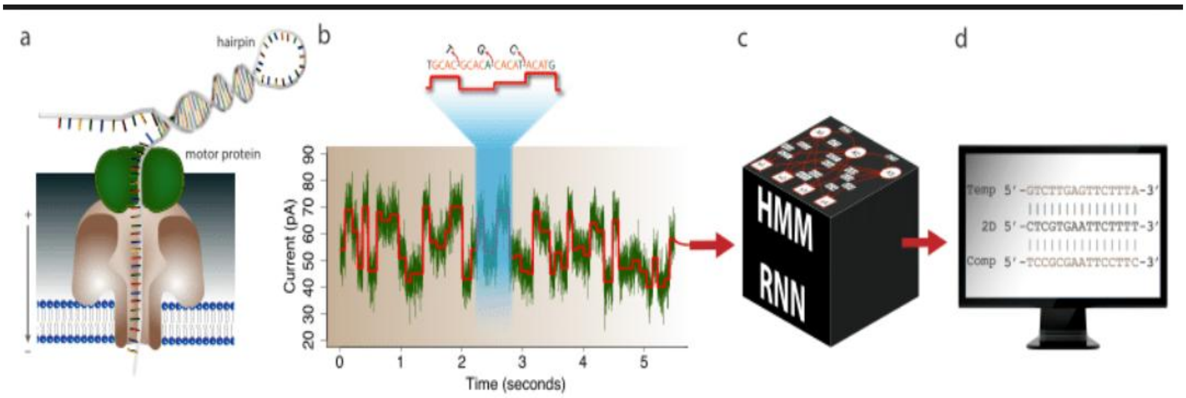
PromethION is a small bench-top system. Nanopore sequencing uses pores formed from proteins, such as haemolysin, a biological protein channel system in *Staphylococcus aureus*

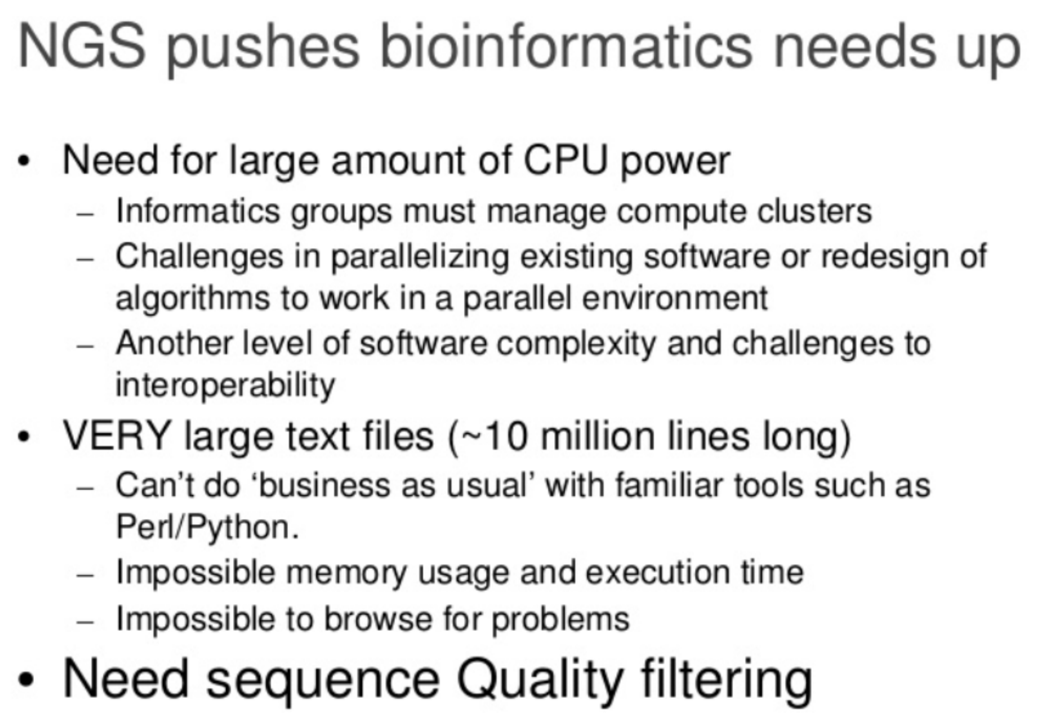
Nanopore is based on the flow of ion current depends on the shape of the molecule translocating through the pore. Since nucleotides have different shapes, each nucleotide is recognized by its effect on the change of the ionic current

Advantage: sample preparation is minimal compared to second-generation sequencing methods, and long read lengths can be obtained in the kbp range. There are no amplification or ligation steps required before sequencing.

Disadvantage: The main problem with this technology is the requirement to optimize the speed of DNA translocation through the nanopore to ensure reliable measurement of the ionic current changes and reduce the high error

Workflow for Nanopore DNA/RNA sequencing:





NGS pushes bioinformatics needs up

- Need for large amount of CPU power
  - Informatics groups must manage compute clusters
  - Challenges in parallelizing existing software or redesign of algorithms to work in a parallel environment
  - Another level of software complexity and challenges to interoperability
- VERY large text files (~10 million lines long)
  - Can't do 'business as usual' with familiar tools such as Perl/Python.
  - Impossible memory usage and execution time
  - Impossible to browse for problems
- Need sequence Quality filtering

APPLICATIONS OF NGS:

DNA

- whole genome sequencing
- Targeted re-sequencing
- ChIP-Seq

RNA

- mRNA
- Whole transcriptome
- Small RNA

These applications can be applied to understand the genomes of existing organisms ,conserving species, understanding the genomics of various disease, finding new mutations associated to them and new diagnostic marker/s.

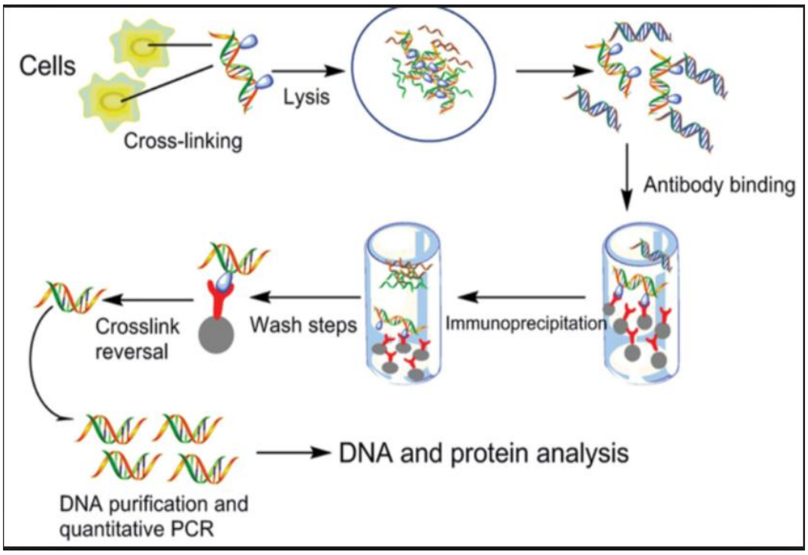
Chromatin Immunoprecipitation

Chromatin immunoprecipitation, or ChIP, refers to a procedure used to determine whether a given protein binds to or is localized to a specific DNA sequence in vivo.

Chromatin immunoprecipitation (ChIP) is a powerful technique for analyzing histone modifications as well as binding sites for proteins that bind either directly or indirectly to DNA.

The technique requires a high- grade Antibody against the protein of interest, which can be used to pull down the bound DNA

Helps in understanding various gene expression regulatory mechanism.



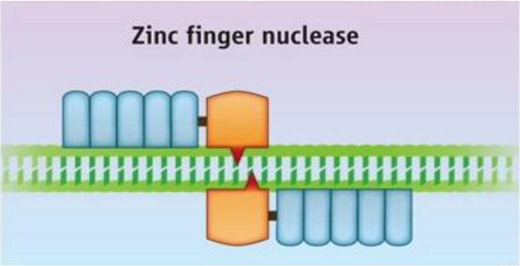
ChIP when coupled with microarray or NGS can help in assessment of chromosomal binding sites for transcription factors or the location of histone modifications at a genomic scale.

GENOME EDITING

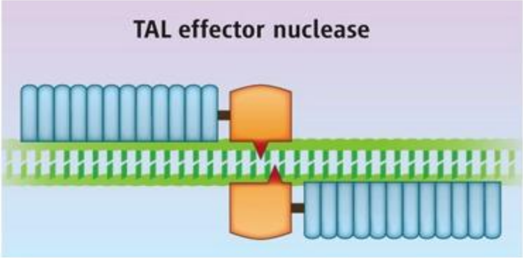
Genome editing tools are sequence-specific nucleases

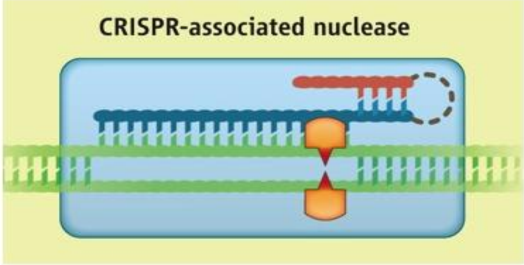
Genome editing tools have two features:

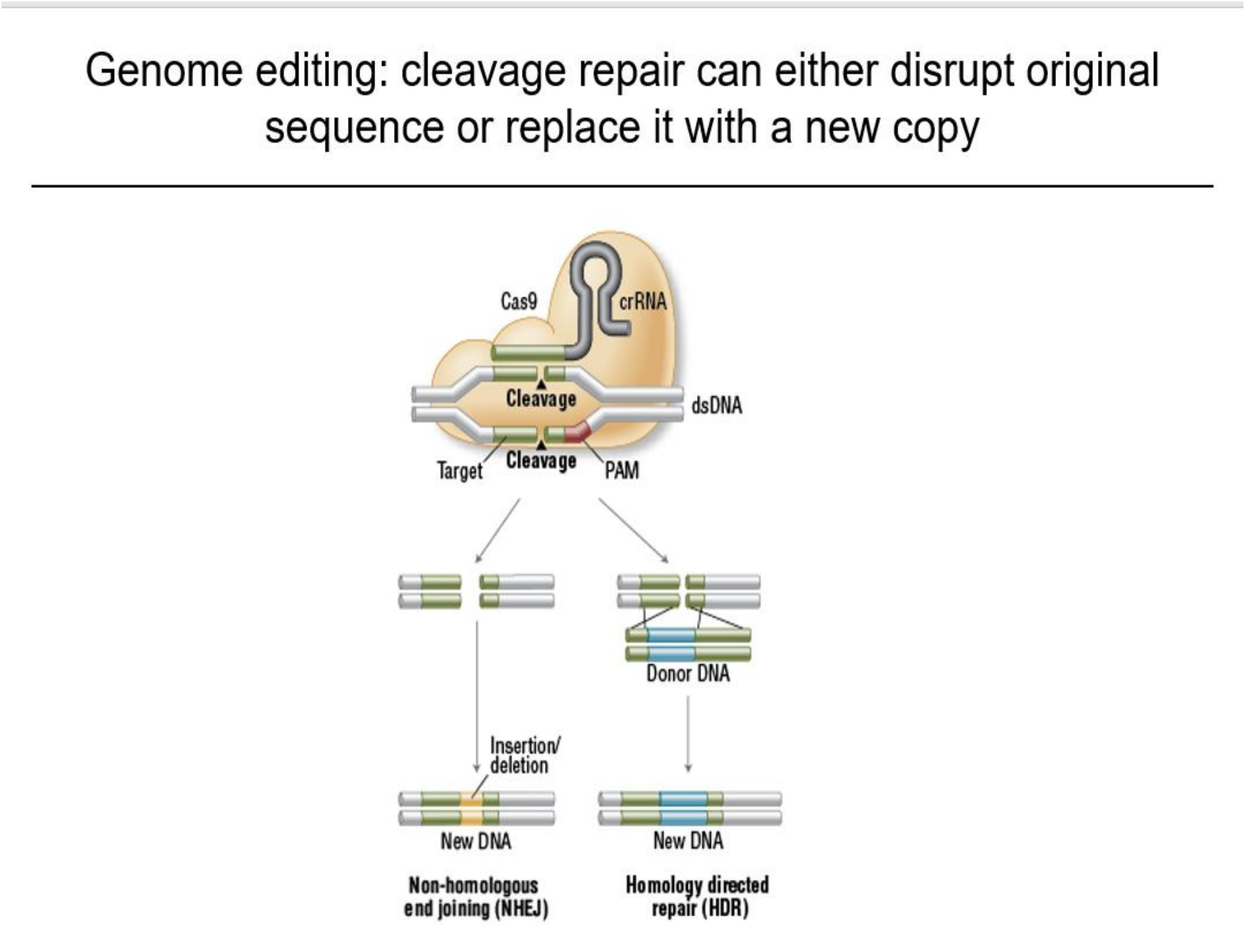
1) Recognize specific DNA sequences (i.e. specific genes or non-coding elements)
2) Cut DNA ("nuclease"), then a scar is left behind.



Transcription activator-like *effector nucleases* (TALEN) are restriction enzymes that can be engineered to cut specific sequences of DNA. They are made by fusing a *TAL effector* DNA-binding domain to a DNA cleavage domain (a *nuclease* which cuts DNA strands).





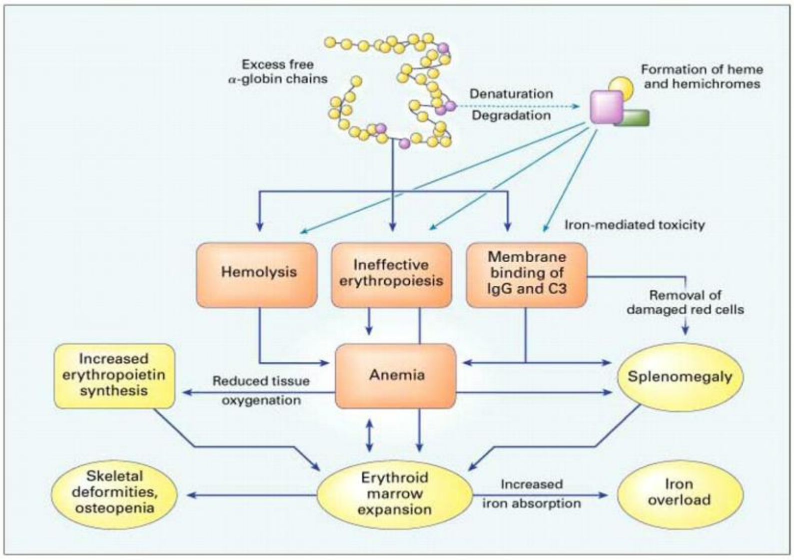


Genome editing: cleavage repair can either disrupt original sequence or replace it with a new copy

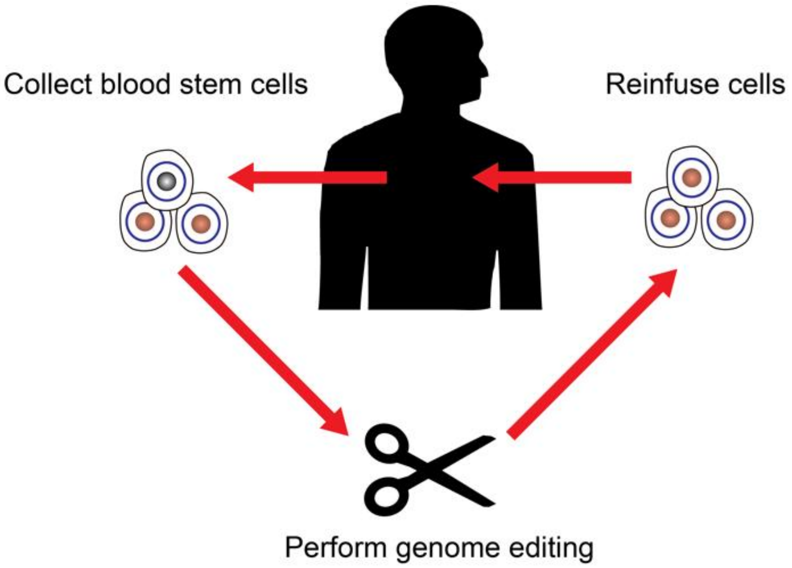
CRISPR is a family of DNA sequences found within the genomes of prokaryotic organisms such as bacteria and archaea.These sequences are derived from DNA fragments from viruses that have previously infected the prokaryote and are used to detect and destroy DNA from similar viruses during subsequent infections. Hence these sequences play a key role in the antiviral defense system of prokaryotes

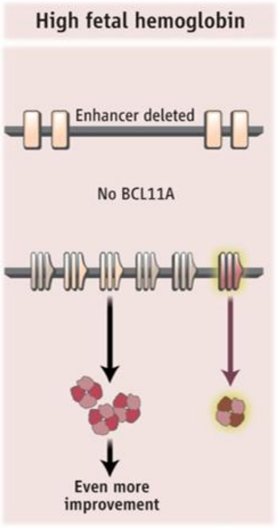
Cas9 (or "CRISPR-associated protein 9") is an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific strands of DNA that are complementary to the CRISPR sequence. Cas9 enzymes together with CRISPR sequences form the basis of a technology known as CRISPR-Cas9 that can be used to edit genes within organisms

The problem in b-thalassemia is too much a-globin relative to b- globin



- Collect blood stem cells from patient with β-thalassemia
- Introduce sequence-specific nucleases to disrupt BCL11A enhancer
- Reinfuse modified blood stem cells to patient





b-thalassemia results from mutations in b-globin, a single gene within a large genome

Gene addition adds a copy of b-globin by semi-random integration into the genome

Currently being tested in early-phase clinical trials

Challenges include: durable high-level expression; ensuring other important genes are not disrupted due to integration

Genome editing offers the promise of precise and permanent genome modification to mimic protective genetic variation (e.g. at BCL11A) or to repair b-globin

Challenges include: effective delivery of genome editing tools to cells to achieve efficient target disruption/repair; ensuring modification is limited to intended target.

Genome editing can be used:

- **For research:** Genome editing can be used to change the DNA in cells or organisms to understand their biology and how they work.
- **To treat disease:** Genome editing has been used to modify human blood cells that are then put back into the body to treat conditions including leukaemia and AIDS. It could also potentially be used to treat other infections (such as MRSA[?]) and simple genetic conditions (such as muscular dystrophy and haemophilia).
- **For biotechnology[?]:** Genome editing has been used in agriculture to genetically modify crops to improve their yields and resistance to disease and drought, as well as to genetically modify cattle that don’t have horns.

SUMMARY:

1. The genome broadly refers to the total amount of DNA of a single cell

(Haploid cell in the case of a diploid organism)

2. Genomics is the study of whole genomes of organisms and incorporates elements from genetics.

3. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the structure and function of genomes.

4. In 1977, two separate methods for the large-scale sequencing of DNA were devised, the chemical cleavage method by Maxam-Gilbert and the Chain termination method by Frederick Sanger.

5. SAGE, a technique developed around 1995, which allows rapid and detailed analysis of overall gene expression patterns.

6. Similary around 1995, the microarray was developed. In DNA microarrays (also known as DNA chips or biochips), each spot contains predefined short (25- to 70-nucleotide) single-stranded DNA oligonucleotides or larger (200- to 800-basepair) double-stranded DNA, which are known as probes. A microarray experiment consists of the hybridization of a sample of target DNA in solution, to large number of probes fixed to the substrate and the difference in expression in measured mostly on the basis of fluorescent tag.

7. Next generation sequencing (NGS), massively parallel or deep sequencing are related terms that describe a DNA sequencing technology which has revolutionised genomic research. Using NGS an entire human genome can be sequenced within a single day compared to the Sanger sequencing method, which took almost more than a decade. NGS can be used to sequence entire genomes or constrained to specific areas of interest, including all 22 000 coding genes (a whole exome) or small numbers of individual genes. There are several sequencers based on various technology.

8. Chromatin immunoprecipitation, or ChIP, refers to a procedure used to determine whether a given protein binds to or is localized to a specific DNA sequence in vivo. The DNA purified using an antibody against the protein bound to DNA can be analyzed using microarray and NGS.

9. Genome editing is the deliberate alteration of a selected DNA sequence in a living cell. A strand of DNA is cut at a specific point and naturally existing cellular repair mechanisms, then fix the broken DNA strands.

10. Genome editing can be achieved using various techniques like Zincfingernucleases, TALENs and CRISPR associated nucleases.

11. The term CRISPR/Cas9 stands for Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR associated protein 9. CRISPR/Cas9 is a system found in bacteria and involved in immune defence. Once the Cas9 scissors cut the DNA just where we intend with the help of a fguide RNA, the cell will try to repair the break using any available DNA

END OF MODULE

QUARANT III:

1. https://pdfs.semanticscholar.org/7669/0918d1ac6df349d6ffb0cf8ccc066d2f126d.pdf
A review of DNA sequencing techniques.

2.Review of massively parallel DNA sequencing technologies
Hugo J. 2011 Dec; 5(1-4): 1–12.

3.Annual Review of Genomics and Human Genetics Volume 9, 2008 Mardis, pp 387-402

Review on Next generation sequencing technologies.

4. Curr Opin Virol. 2019 Aug 23;38:81-88. doi: 10.1016/j.coviro.2019.07.001. [Epub ahead of print]

Elimination of infectious HIV DNA by CRISPR-Cas9.

5. Cell Volume 157, ISSUE 6, P1262-1278, June 05, 2014. Development and Applications of CRISPR-Cas9 for Genome Engineering

6. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology Clin Microbiol Rev. 2009 Oct; 22(4): 611–633.

7. DNA Microarray Technology: Devices, Systems, and Applications. Annual Review of Biomedical Engineering Vol. 4:129-153 (Volume publication date August 2002)

8. The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology* volume 20, pages490–507 (2019)

9. Use of serial analysis of gene expression (SAGE) technologyJournal of Immunological Methods 250 (2001) 45–66

10. Use of serial analysis of gene expression (SAGE) technology. J Immunol Methods. 2001 Apr;250(1-2):45-66.

QUADRANT 1V:

QUESTIONS

Multiple choice Questions

1. Which of the following is incorrect about a microarray?
   a) It is a slide attached with a high-density array of immobilized DNA oligomers representing the entire genome of the species under study
   b) Array of immobilized DNA oligomers cannot be cDNAs
   c) Each oligomer is spotted on the slide and serves as a probe for binding to a unique complementary cDNA
   d) It is the most commonly used global gene expression profiling method

2. Which of the following is untrue about the drawbacks of SAGE?
   a) One or two sequencing errors in the tag sequence can lead to ambiguous or erroneous tag identification
   b) Correctly sequenced SAGE tag sometimes may correspond to several genes or no gene at all
   c) Correctly sequenced SAGE tag always corresponds to several genes
   d) The drawback with this approach is the sensitivity to sequencing errors

3. Chain-termination is a type of _______________
   a) Sequencing
   b) Vector generation
   c) Antibiotic production
   d) Gene manipulation

4. The first significant DNA sequence to be obtained was that of ________
   a) Lambda
   b) Plasmid
   c) Lactose
   d) Mammals

5. CRISPR refers to repeated sequences located in the

   a) Bacterial DNA
   b) Fungal DNA
   c) Viral DNA
   d) Viral RNA

Answer Key

Q1. b

Q2. C

Q3. a

Q4. a

Q5. a

SAQ(3 marks each):

1. Write two main difference between Chemical Cleavage and Chain Termination method of sequencing.
2. What is the full form of CRISPR? How does the system work ?
3. Write the steps involved in SAGE.
4. Write 6 applications for Next generation sequencing.
5. How can CRISPR be used to cure thalassemia

LAQ(5 marks each)

1. What is Human Genome project? What did we learn from this project?
2. Mention one third generation NGS system? Explain the technology. What are it's advantages and disadvantages
3. What is Illumina sequencing? Explain how does the platform work.
4. Write five applications of microarray.
5. Explain the term sequencing by ligation and sequencing by synthesis.