Elective Paper.

Unit 2.

01] History of Sequencing

→ Foundations of DNA Sequencing 1871-1952

① The history of DNA sequencing can be traced backs to the late 19th century when scientist like Friedrich Miescher laid the groundwork in cell chemistry & nucleic substances. By 1929, the chemical composition of nucleic acids & the structure of DNA's backbone were understood, revealing the existence of nitrogen containing bases (A, C, T, G) & alternating sugar phosphate backbone.

The Revolutionary Discovery 1944-1952

In 1944, Oswald Avery, Madyn McCarty's groundbreaking research demonstrated that DNA, not protein carries genetic information & can transform cellular properties. This discovery led to significant shift in scientific focus towards DNA research.

In 1952, Erwin chargaff & Linus Pauling made important contributions by revealing crucial chemical properties of DNA, including the base pairing rules (A=T & C=G)

The Sanger Method (1965-1982)

The first sequencing methods were introduced in 1968, with a heavy focus on RNA sequencing due to its relative simplicity. It was not until 1979 that Frederick Sanger developed a method for DNA sequencing that became the gold standard for over a decade. Sanger's method enabled a major milestone, the sequencing of the first whole genome of a bacteriophage in 1982.

Second generation Sequencing (1996-2006)

The emergence of second generation sequencing technologies began with the introduction of pyrosequencing in 1996. These methods including Roche/454, SOLiD & Illumina allowed for high throughput sequencing & significantly reduced the associated cost. Roche/454 life sciences released the first high throughput sequencer in 2005, followed by Illumina's Genome Analyzer in 2006, propelling the field of genomics into new era.

Third Generation Sequencing (2008)

Third generation sequencing technologies revolutionized DNA sequencing by providing long read capabilities. Pacific Biosciences SMRT sequencing uses fluorescent nucleotides to generate long reads. & oxford Nanopore Technologies nanopore technology measures disruptions in electrical current as individual bases pass through a nanopore. These technologies have facilitated genome assembly & structural analysis.

Maxam Gilbert

DNA sequencing refers to methods for determining the order of the nucleotide bases adenine, guanine, cytosine & thymine in a molecule of DNA

Two main methods are widely known to be used to sequence DNA

1. The chemical Method (also called Maxam-Gilbert method after its inventors)
2. The chain Termination Method (also known as Sanger dideoxy method)

Maxam gilbert technique depends on the relative chemical lability of different nucleotide bonds, whereas the sanger

through elongation of DNA sequence by incorporating dideoxynucleotide into the sequence.

The chain termination method is the method more usually used because of its speed & simplicity.

(1) In 1976-1977, Allan Maxam & Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA & subsequent cleavage at specific bases

(2) The method required radioactive labelling at one end & purification of the DNA fragment to be sequenced.

(3) Chemical treatment generated breaks at a small proportions of one or two of the four nucleotide based in each of four reactions (G, A+G, C, C+T)

(4) Thus a series of labelled fragments is generated from the radiolabelled end to the first 'cut' site in each molecule

(5) The fragments in the four reactions are arranged side by side in gel electrophoresis for size separation

(6) To visualize the fragment the gel is exposed to xray film for autoradiography yielding a series of dark bands each corresponding to a radiolabelled DNA fragment from which the sequence may be inferred

II Sanger Sequencing

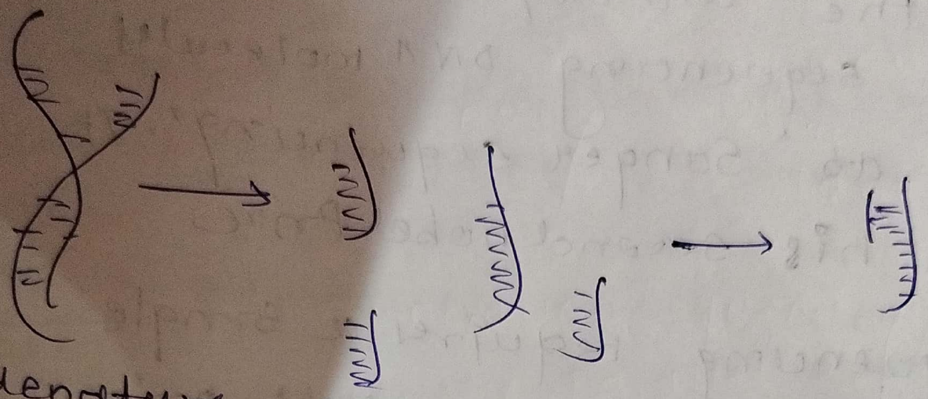
(1) The first major breakthrough in sequencing technology was made by Fredrick Sanger in 1977, when he & his colleagues introduced the "dideoxy chain termination method for sequencing DNA molecule also known as 'Sanger sequencing'. It earned him his second nobel Prize

(2) Sanger sequencing requires a single stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotide triphosphate (dNTPs) and modified dideoxynucleotide triphosphate (ddNTPs) the latter of which terminate DNA strand elongation. The DNA sample is divided into four separate sequencing reactions & to each reaction only one of the four dideoxynucleotides are added. (A, T, G or C)

(3) The dideoxynucleotides are added at very low concentrations, such that they are very rarely incorporated

(4) Because of this, many different DNA strand lengths are formed, each with radioactive nucleotide at their terminus.

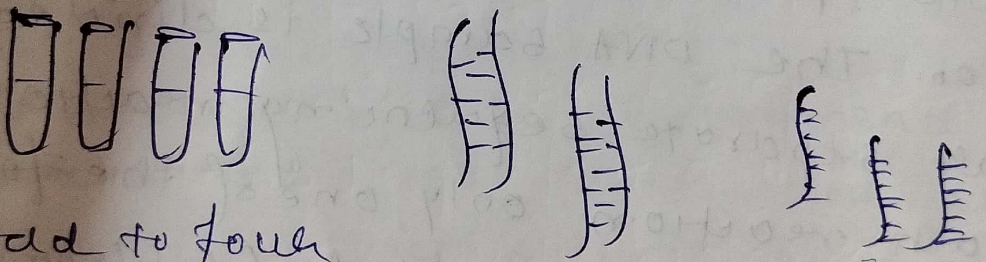
By 'lining up' all the varied length strands from all four reactions, one is able to see where each nucleotide occurs.



denature dsDNA using heat

make multiple copies of a segment

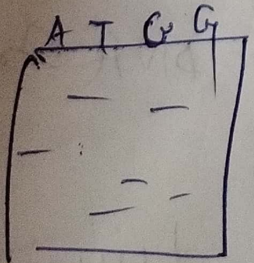
attach primer



Add to four polymerase solutions

grow complementary chains until termination dye

denature the grow chains



Electrophoresis this four soln

However Sanger Sequencing lacked Automation & was extremely time consuming.

Q3] -

First Generation	Second Generation
(1) A low throughput method used to determine a portion of nucleotide seq of an individual genome	(1) A high-throughput method used to determine a portion of the nucleotide seq of an individual genome
(2) Dideoxy chain termination method	(2) Massive parallel Sequencing.
(3) First commercialized by Applied Biosystem	(3) Dominant platform is Illumina
(4) Can only process a single DNA fragment at a time	(4) Process millions of fragments simultaneously at a time
(5) A less sensitive method	(5) Highly sensitive
(6) limit detection is 5-20	(6) limit detection is less than 1%.

(7) cost effective
Fast
up to 20
sample

(8) Time consuming
& costeffective
upto 20 sample

(9) Gold standard
for clinical
research sequencing

(10) Becoming
common in
clinical
labs

4 a) NGS (Next Generation Sequencing)

NGS is massively parallel sequencing technology that offers ultra high throughput, scalability & speed. The technology is used to determine the order of nucleotide in entire genome or targeted regions of DNA or RNA. NGS has revolutionized the biological sciences allowing labs to perform a wide variety of applications & study biological system at a level never before possible.

Applications of NGS

(1) Rapidly sequence whole genome
(2) Deeply sequence target regions
(3) utilize RNA sequencing (RNA seq) to discover novel RNA variants & splice site, or quantify mRNAs for

gene expression analysis.

(4) Sequence cancer samples to study rare somatic variants, tumor subclones,

(5) study the human microbiome

(6) Identify novel pathogens

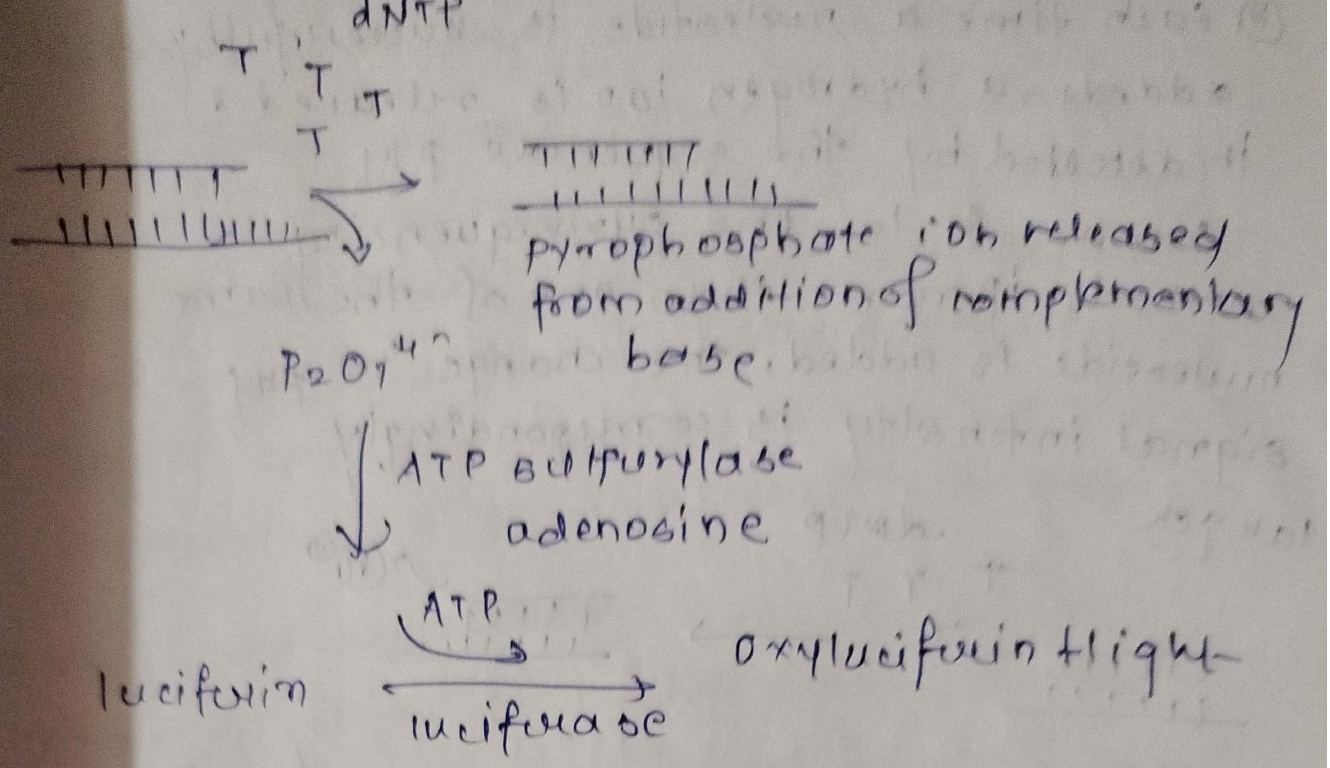
Q] Roche 454 Pyrosequencing.

(1) Pyrosequencing is based on the sequencing by synthesis principle where a complementary strand is synthesized in the presence of polymerase enzyme.

(2) In contrast to using dideoxynucleotides to terminate chain amplification (as in Sanger sequencing), pyrosequencing instead detects the release of pyrophosphate when nucleotides are added to DNA chain.

(3) It initially uses the emulsion PCR technique to construct the polonies required for sequencing & removes the complementary strand. Next a SS DNA sequencing primer hybridizes to the end of the strand. then the four different dNTPs are then sequentially made to flow in & out of the wells over the polonies.

(4) When the correct dNTP is enzymatically incorporated into the strand, it causes release of pyrophosphate.
In the presence of ATP sulfurylase & adenosine, the pyrophosphate is converted into ATP.
(5) This ATP molecule is used for luciferase catalyzed conversion of luciferin to oxyluciferin which produces light can be detected with a camera.
(6) The relative intensity of light is proportional to the amt of base added
Ie a peak of twice the intensity indicates two identical bases have been added in succession)
However, the method was eclipsed by other technologies & in 2013 454 pioneer Roche announced the closure of 454 life sciences & the discontinuation of the 454 pyrosequencing platform.

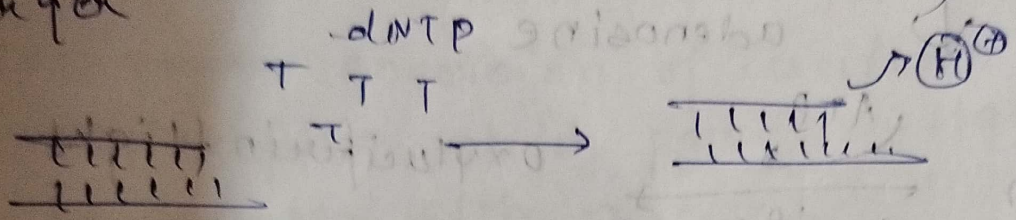


⑥ Ion torrent Semiconductor Sequencing

① Ion torrent sequencing uses a sequencing by synthesis approach, in which a new DNA strand, complementary to the target strand, is synthesized one base at a time. A semiconductor chip detects the hydrogen ions produced during DNA polymerization.

② Following polony formation using emulsion PCR, the DNA library fragment is flooded sequentially with each nucleoside triphosphate (dNTP) as in pyrosequencing. The dNTP is then incorporated into new strand if complementary to the nucleotide on the target strand.

added, a hydrogen ion is released & it detected by the sequencer's pH sensor. As in the pyrosequencing method, if more than one of the same nucleotide is added, the change in pH signal intensity is correspondingly larger



④ Ion torrent sequencing is the first commercial technique not to use fluorescence & camera scanning

It is faster & cheaper

Dis It can be difficult to enumerate the number of identical bases added consecutively

⑤ Sequencing by ligation (SOLiD)

① SOLiD is an enzymatic method of sequencing that uses DNA ligase, an enzyme used widely in biotechnology for its ability to ligate double-stranded DNA strands

② Emulsion PCR is used to immobilise a ss DNA primer binding region

(known as adapter) which has been conjugated to the target sequence (i.e the seq that is to be sequenced) on a bead. These beads are then deposited onto a glass surface. a high density of beads can be achieved which in turn increases the throughput of technique

(3) Once bead deposition has occured a primer of length N is hybridized to the adapter, then the beads are exposed to a library of 6-mer probes which have different fluorescent dye at the 5' end & a hydroxyl grp at the 3' end.

(4) Only a complementary probe will hybridize to the target seq, adjacent to the primer.

(5) A phosphorothioate linkage between bases 5 & 6 allows the fluorescent dye to be cleaved from the fragment using silver ions. This cleavage allows fluorescence to be measured (four different fluorescent dyes are used, all of which have different emission spectra) and also generate a 5' phosphate grp which can undergo further ligation. Once the first round of

sequencing is completed, the extension product is ~~method~~ melted off and then a second round of sequencing is performed with a primer of length N-1. Many rounds of sequencing ~~is performed~~ using shorter primers each time (ie N-2, N-3 etc) & measuring the fluorescence ensures that the target is sequenced

(4) Due to the two base sequencing method (since each base is effectively sequenced twice) the SOLiD technique is highly accurate (at 99.999%) with a sixth primer it is the most accurate of the second generation platforms) and also inexpensive It can complete a single run in 7 days and in that ~~time~~ can produce 30 Gb of data.

(5) Unfortunately its main disadvantage is that read lengths are short, making it unsuitable for many applications

Advantages of Next generation

(1) Faster turnaround time for samples with high volume

(2) Higher productivity with sample multiplexing

(3) It can sequence thousands of samples simultaneously

(4) With the same amt of DNA input, more info is created

Disadvantages of Next generation sequencing

(1) Next generation sequencing can reveal a variety of molecular abnormalities. The clinical relevance of many of the anomaly discovered is yet unknown

(2) As so many genes are being evaluated, there could be unexpected results, such as risk factors for other diseases or unidentified variants

(3) Sequencing a small number of targets like 1-20 targets is time consuming

(i) A new cohort of techniques has since been developed using single molecule sequencing and single real time sequencing removing the need for clonal amplification.

(ii) This reduces errors caused by PCR, simplifies library preparation and most importantly gives a much higher read length using higher throughput platforms. Examples Pacebioscience platform which uses SMRT Sequencing to give read lengths of around one thousand bases & Helicos Biosciences which utilises single molecule sequencing & therefore does not require amplification prior to sequencing.

(iii) Oxford Nanopore Technologies are currently developing silicon based nanopores which are subjected to a current that changes as DNA passes through the pore. This is anticipated to be a high throughput rapid method of DNA sequencing, although problems such as slowing transportation through the pore must be addressed.

PacBio

1. In 2011, PacBio released Single molecule Sequencing in Real Time (SMRT) a new type of sequencing technique
2. Unlike previous techniques, SMRT could sequence individual molecules and individual cells for the first time
3. It was used to sequence the E.coli genome with 99.999% accuracy.
4. SMRT is a third generation sequencing technology developed by Pacific Biosciences (PacBio)
5. Between 2011 & 2018 the technology evolved from being able to read fragments of around 1,000 bases to 30,000 bases in one go known as long read

comparatively, earlier techniques like 454 Sequencing could only read up to 400 bases.

Initially, its speed & long read ability meant lower accuracy than early methods However because it is so fast and any errors are random, it's possible to repeat experiment & achieve upto to 99.99% accuracy.

How does SMRT work

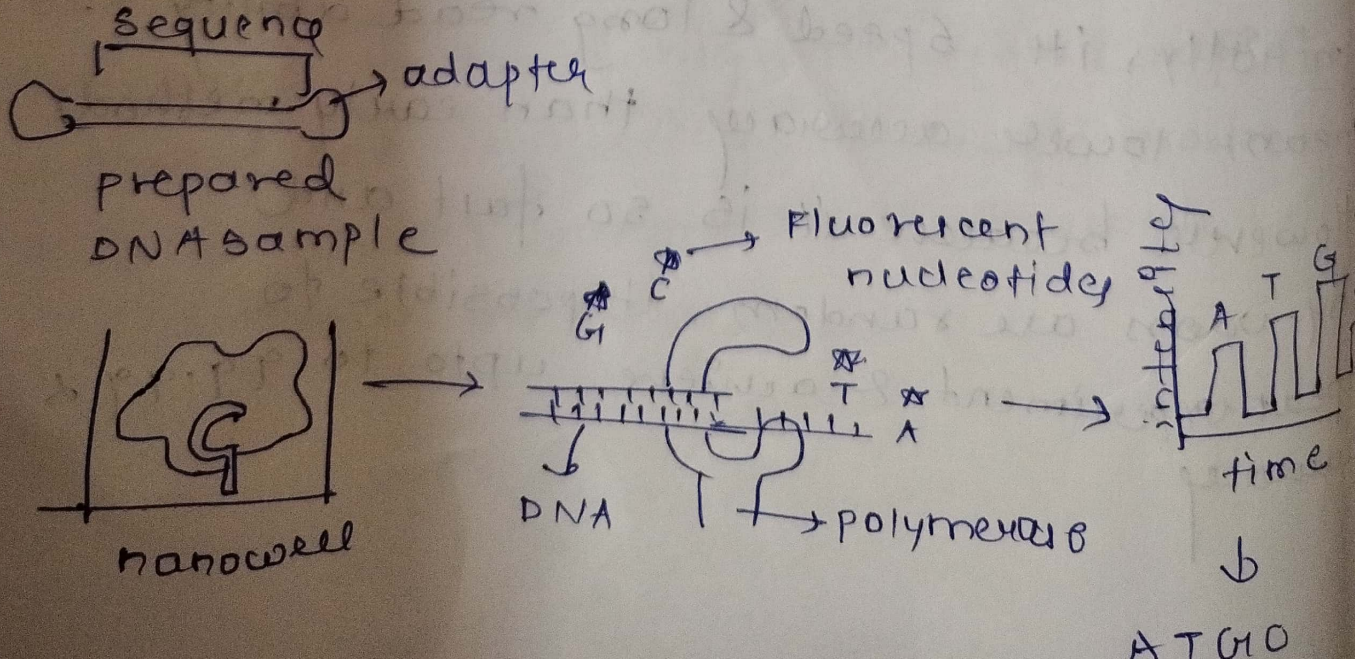
(1) SMRT involves a single stranded molecule DNA, bound to a DNA polymerase enzyme

(2) The bound pair enter a sequencing chamber called a flow cell

(3) Like in Sanger sequencing, the DNA polymerase adds complementary, fluorescently labelled bases to the DNA strand

(4) As each labelled base is added, the fluores colour of the base is recorded before the fluorescent label is cut off. The next base in the DNA chain can the be added & recorded

(5) This is translated into a sequence by an algorithm.



Single Molecule Sequencing in Real Time

A prepare DNA sample is bound to the polymerase via an adapter & flow onto the flow cell.

A DNA polymerase is connected to the bottom of each nanowell & a mix of fluorescently labelled nucleotides is added.

The incorporation of each fluorescent nucleotide leads to a burst of light captured in the raw video data

Benefits

1. cost & speed
2. Sensitivity
3. length
4. Accuracy.

Oxford Nanopore Technology

(1) In 2015, the first Oxford Nanopore Technology (ONT) sequencing device was released an entirely new way to sequence DNA

(2) unlike previous techniques, which were based on DNA replication, ONT doesn't use any DNA polymerases. Instead, it's based on a small barrel shaped protein in a membrane called nanopore, & measures changes to electrical current

(3) It was used in 2015 during an Ebola outbreak to help with genomic surveillance

(4) Oxford Nanopore technology is a type of third generation sequencing technology. It doesnot use any DNA polymerase

(8) Instead, it uses a barrel shaped protein called a hemolysin - naturally found as a pore in a cell membrane, regulating which molecules can enter or leave cell this is called nanopore

(9) α-Hemolysin has a diameter of one nanometer just big enough to allow a single strand of DNA through.

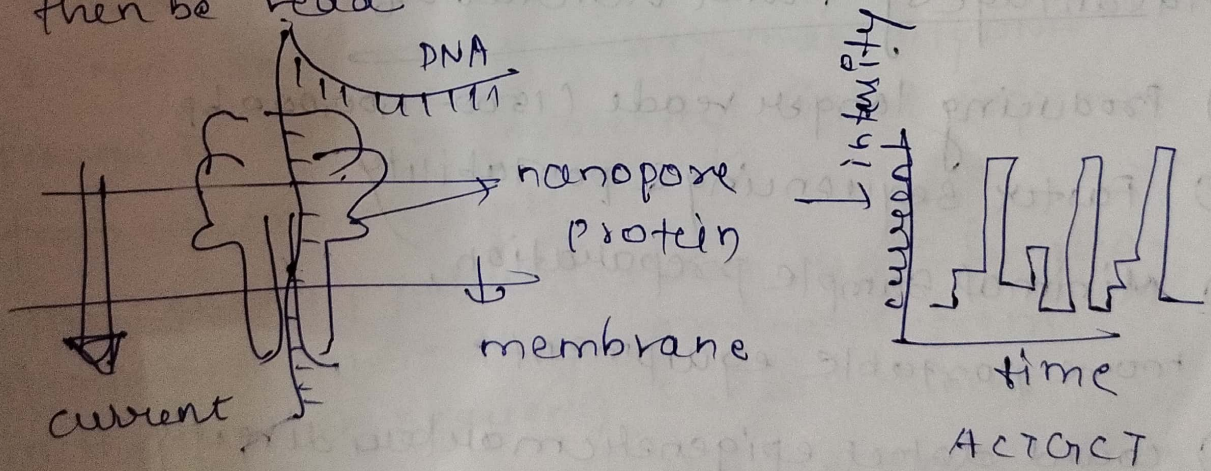
How does ONT works

In ONT the α-hemolysin nanopore is embedded into an artificial membrane inside a sequencing chamber. When a current is applied to the membrane, the DNA travels through the nanopore

(2) As the DNA travels through the nanopore it obstructs the current flowing across the membrane

(3) The four base of the DNA (A, T, C, G) are of different shape & size so cause variations in current

(4) These variations are measured by an electronic chip. An algorithm converts the data into seq which can then be read



Nanopore proteins are embedded into an artificial membrane inside the sequencing flowcell.

The obstruction of a nanopore by a DNA fragment ~~reads~~ to change in the reads current that is measured continuously by an electronic chip integrated within the flowcell

Benefits

1. Can generate ultra long reads
2. Instrument is very small between size of phone & a microwave
   It make technology more accessible requiring only laptop to run it

Advantage of third Generation

1. Producing longer reads (1500-1,00,000 bp)
2. Faster sequencing & portability.
3. Minimal sample preparation, more portable equipment
4. could detect epigenetic markers directly as they produce unique signal

Disadvantages

① Error rate remain notably higher due to instability of the molecular machinery involved

② rapid pace of sequencing can cause signals from adjacent base to blur, complicating the interpretation of signals & the analysing of sequence

These error hinder the accurate classification of individual genetic difference within same species.

9) Hi-C

① Hi-C to comprehensively detect chromatin interactions in the mammalian nucleus.
This method is based on chromosome conformation capture, in which chromatin is crosslinked with formaldehyde, then digested & re-ligated in such a way that only DNA fragments that are covalently linked together from ligation products.

② The ligation product contain the information of not only where they originated from in the genomic sequence but also where they reside, physically in the 3D organization of the genome.

(4) In Hi-C a biotin labeled nucleotide is incorporated at the ligation junction, enabling selective purification of chimeric DNA ligation junctions followed by deep sequencing

(5) The compatibility of Hi-C with next generation sequencing platforms makes it possible to detect chromatic interactions on an unprecendented scale

(6) This advance gives Hi-C the power of both explore the biophysical properties of chromatin as well as the implications of chromatin structure for the biological functions of the nucleus

(7) A massively parallel survey of chromatin interaction provides the previously missing dimension of spatial context to other genomic studies

(8) This spatial context will provide a new perspective to studies ochromatin & its role in genomic regulation in normal conditions & in disease

Pros

① Allows detection of long range DNA interactions

② High throughput method

Cons

① Detection may result from random chromosomal collisions

② Less than 1% of DNA fragments actually yield ligation products

③ Due to multiple steps, the method requires large amt of starting material

⑥ ChIP Seq :—

① By combining chromatin immunoprecipitation (ChIP) assays with sequencing, ChIP Sequencing (ChIP-Seq) is a powerful method for identifying genome-wide DNA binding sites for transcription factors & other proteins.

② By ChIP protocols, DNA bound protein is immunoprecipitated using a specific antibody. The bound DNA is then coprecipitated, purified & Sequenced

The application of next generation sequencing (NGS) to chip has revealed insight into gene regulation events that play a role in various diseases & biological pathways, such as development & cancer progression. chip seq enables thorough examination of the interactions between proteins & nucleic acids on a genome wide scale

Advantages of chipseq

1. captures DNA targets for transcription factors or histone modifications across the entire genome of an organism
2. Defines transcription factor binding sites
3. Reveals gene regulatory networks in combination with RNA sequencing & methylation analysis

How does chip seq works

1. chip seq identifies the binding sites of DNA associated proteins and can be used to map global binding sites for a

of DNA protein complexes.

(3) Samples are then fragmented & treated with an exonuclease to trim unbound oligonucleotides.

(4) Protein specific antibodies are used to immunoprecipitate the DNA protein complex. The DNA is extracted & sequenced giving high resolution sequences of protein binding site

6] Metagenomics :-

(1) Metagenomics is a new approach to study microorganisms obtained from specific environment by functional gene screening or sequencing analysis

(2) Metagenomic studies focus on microbial diversity, community constitute, genetic & evolutionary relationships, functional activities & interactions & relationships with environment.

(3) Sequencing technologies has evolved from shotgun sequencing to high throughput

next generation sequencing (NGS) &
third generation sequencing (TGS).
NGS & TGS have shown the advantage of
rapid detection of pathogenic microorganisms
(4) With the help of new algorithms, we
can better perform the taxonomic profiling
& gene prediction of microbial species

(5) Functional metagenomics is helpful to screen
new bioactive substances and new functional
genes from microorganisms & microbial
metabolits.

(6) Metagenome also known as microbial
environmental genome (eDNA) is defined as
"the genome of the total microbes found
in nature."

(7) Metagenome currently refers to the sum
of the genomes of bacteria & fungi in
environmental samples.

(8) Metagenomics is the study of a collection
of genetic material from a mixed
community of organisms.

(9) The main areas of concern of metagenomics
research are microbial diversity, population
structure, genetic & evolutionary relationship

functional activity, & cooperative relationships & relationship with the environment.

(3) Metagenomics research is developing rapidly in medicine, agriculture, environmental protection & other fields.

Q) DNA Sequencing (DNA-Seq)

(1) DNA sequencing determines the order of the four chemical building blocks called 'bases' that make up DNA molecule

(2) Sequencing DNA means determining the order of the four chemical building blocks that make up DNA molecule. The sequence tells scientists the kind of genetic information that is carried in a particular DNA segment

(3) For eg, scientist can use sequence information to determine which stretches of DNA contain genes & which stretches carry regulatory instructions ; turning genes on or off.

(4) In addition, and importantly, sequence data can highlight changes in a gene that may cause disease

(5) In the DNA double helix, the four chemical bases always bond with the same partner to form "base pairs": Adenine pairs with thymine & cytosine with Guanine.

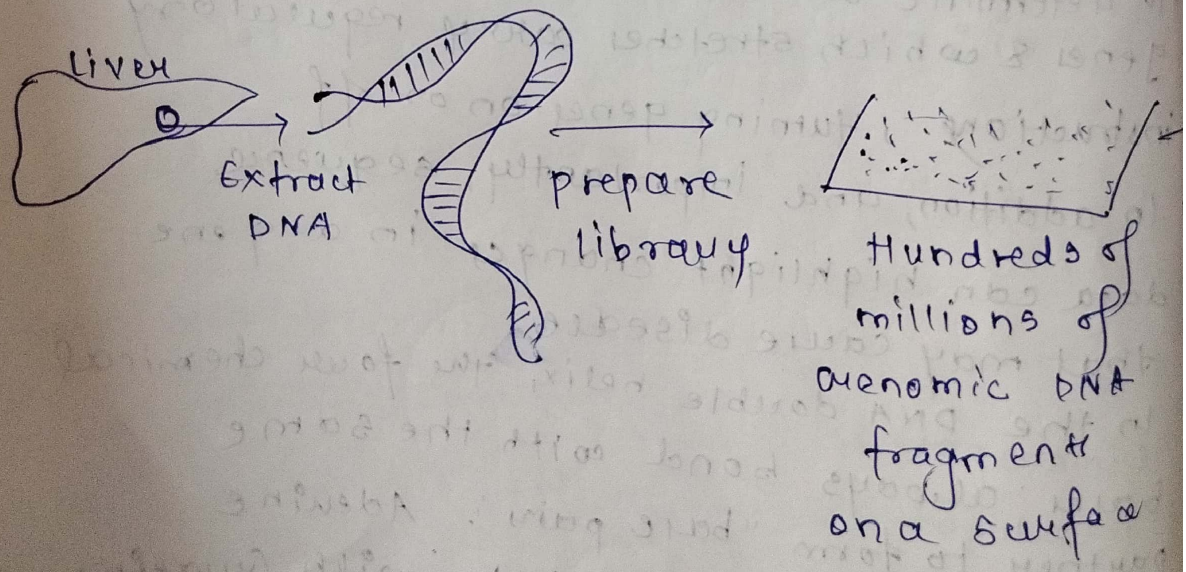
(6) This pairing is the basis for the mechanism by which DNA molecules are copied when cells divide, & the pairing also underlies the methods by which most DNA sequencing experiments are done.

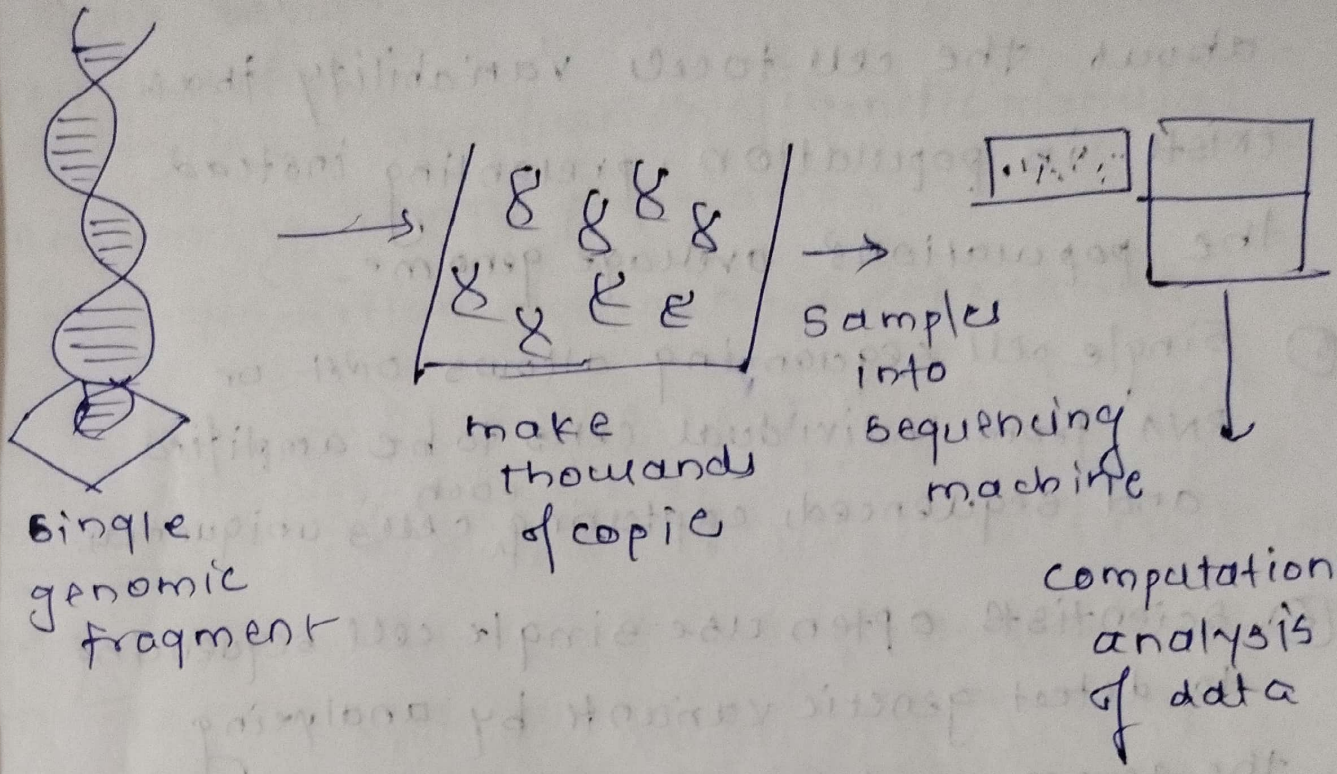
The human genome contains about 3 billion bp that spell out the instructions for making & maintaining a human being.

Application

(1) It can be used to find genes, segments of DNA that code for a specific protein or phenotype

(2) If a region of DNA has been sequenced it can be screened for characteristic features of genes





Single cell Sequencing

① Singe cell sequencing is a collection of methods that researchers use to isolate & analyze sequence information from individual cells.

② single cell sequencing techniques allow researchers to understand more than ever before about cells inner workings.

③ Many traditional sequencing method cannot help researchers to analyze material from individual or small numbers of cells - rather they sequence bulk cell populations where a large number of cells, with their contents of interest, are pooled prior to analysis.

(4) Studying cells in bulk masks information about the cell-to-cell variability that exists in population + presenting instead the population's average genome.

(5) Single cell sequencing allows DNA or RNA from individual cells to be amplified and sequenced, capturing each cell's uniqueness.

(6) Scientists often use single cell sequencing to detect genetic variants by analyzing the genome, understand epigenetic variation by sequencing the methylome, or track gene expression differences by investigating the transcriptome of individual cells in a population.

(7) Through these studies, researchers can identify rare yet important cell subtypes within heterogenous cell populations

(8) Sc RNA has emerged as an important technique for identifying differences in cell that otherwise appear homogenous & understanding cellular responses on the molecular level.

Some basic steps:

1) cell isolation
2) extraction & amplification of genetic material
3) sequencing library preparation.
4) Next generation sequencing & data analysis.

RNA Seq:

(1) RNA seq (RNA sequencing) is a technique that can examine the quantity & sequences of RNA in a sample using next generation sequencing
(2) It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.
(3) RNA seq lets us to investigate and discover the transcriptome, the total cellular content of RNAs including mRNA, rRNA & tRNA

Applications

(1) RNA seq can give vital information about the function of genes.
(2) It also captures information about alternative splicing events which produce different transcripts from one single gene sequence
(3) Helps us to identify post transcriptional modifications that occur during mRNA processing such as polyadenylation & 5' capping

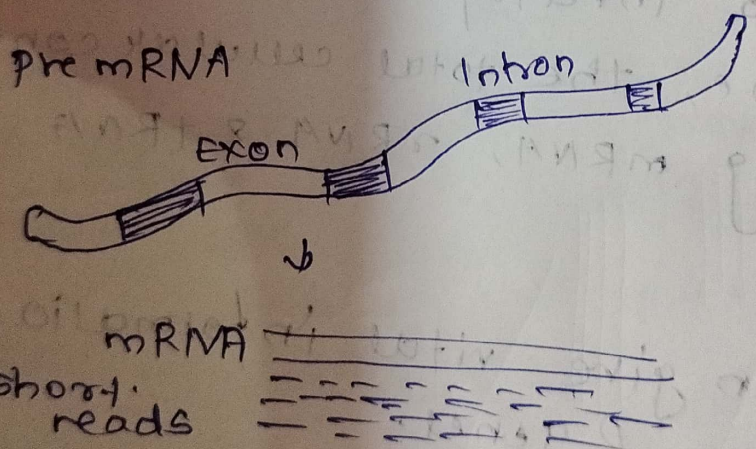
the biology of a cell ... assess changes that may indicate disease

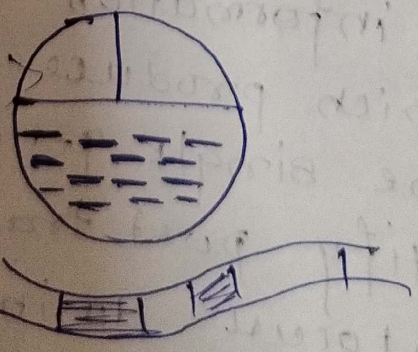
How does RNA seq work

① Early RNA seq technique used Sanger sequencing technology. a technique that

② A RNA workflow has several steps.

1) RNA extraction
2) Reverse transcription into cDNA
3) Adapted ligation
4) Amplification
5) Sequencing





Short read is split by intron when aligning to reference genome

4) Illumina Sequencing :-

(1) Illumina NGS technology utilizes a fundamentally different approach from the classic Sanger chain termination method.

(2) It leverages sequencing by synthesis (SBS) technology tracking the addition of labeled nucleotides as the DNA chain is copied in a massively parallel fashion.

(3) Illumina sequencing process

A) DNA library

Break the genome DNA to form DNA fragments, add adapter at both ends & construct a single stranded DNA library.

After adding DNA library into flowcell, DNA fragments will be attached to the surface of flowcell when passing through flowcell

B) Bridge PCR

Each DNA fragment is clustered in its position.

After bridge amplification each cluster contains many copies of ss DNA template

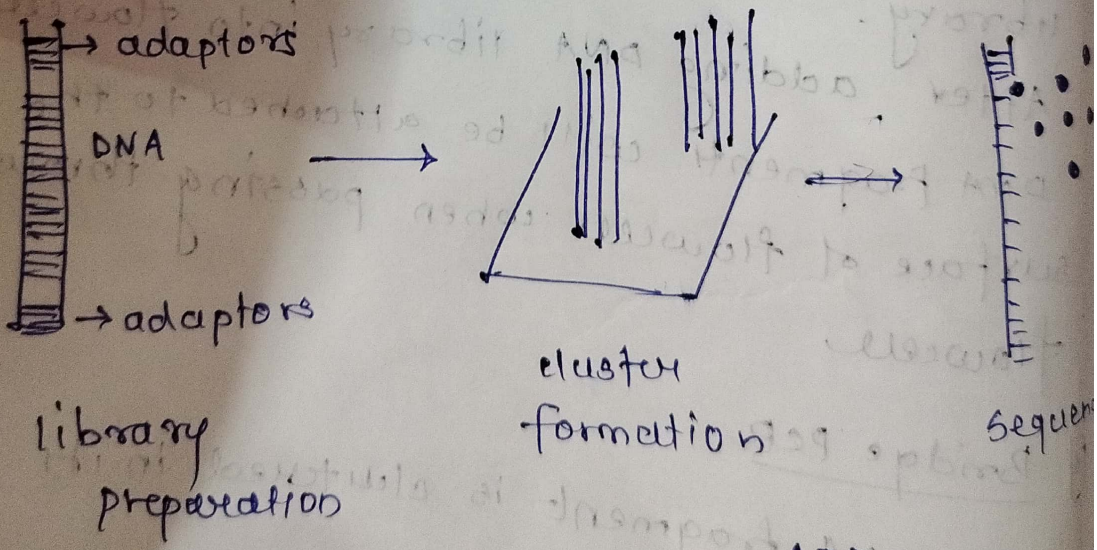
d] SBS

DNA polymerase, primers & four dNTPs with specific fluorescence are added to the reaction system at the same time.

The 3' end of these dNTPs is connected with an azide group, which can block the incorporation of the next base, so only one base can be extended at a time.

After washing off the remaining dNTP & enzyme with water, photos are scanned.

After scanning, add chemical reagents cut off the azide group & quenching fluorescence & next cycle ~~seq~~.



Library preparation

ATP.
TTP.
CTP.
GTP.

TCCA → Sequence identification

a) First Generation Sequencing:-

(1) First Generation Sequencing also known as Sanger Sequencing refers to the chain termination method that was developed by Fredrick Sanger & coworkers in 1977.

(2) The DNA molecule is amplified with modified nucleotides (ddNTPs) and the addition of only one base per cycle is allowed.

(3) The DNA is then amplified with varying length each strand is one basepair longer than the previous molecule. These molecules are then separated by capillary electrophoresis

(4) Due to the fluorescent dye at the end of each fragment, we are able to identify the base (A, G, T or C) by reading the different colour signal

Advantages

1) Long reading sequences. Each assembly in read out (especially for GC rich highly repetitive DNA areas)

2) Smaller depths of sequencing required for good coverage

3) Easy to analyse

4) Relatively data storage required

Disadvantage:-

(1) low sensitivity (high allele frequency of cancer needed)
(2) Scalable to few genes only
(3) unable to detect chromosomal aberration
(4) insensitive to copy number alteration
(5) High cost per base
(6) lower turnaround time

d) COG database:-

(1) The two COGs were constructed by applying the criterion of consistency of genome specific best hits to the results of an exhaustive comparison of all protein sequences from these genomes

(2) The database comprises 2091 COGs that include 56-83% of the gene products from each of the complete bacterial & archaeal genomes & ~35% of those from the yeast Saccharomyces cerevisiae genome

(3) The COG database is accompanied by the COGNITOR program that is

used to fit new proteins into the COGs and can be applied to functional & phylogenetic annotations of newly sequenced genomes

(4) The availability of multiple, essentially complete genome sequences of prokaryotes & eukaryotes spurred both the demand and the opportunity for the construction of an evolutionary classification of genes from these genomes.

(5) Such a classification system based on orthology relationships between genes appears to be a natural framework for comparative genomics & should facilitate both functional annotation of genomes & large scale evolutionary studies

9) VirGen database :-

(1) VirGen is a comprehensive viral genome resource that organizes the sequence space of viral genomes in a structured fashion. It has been developed with the objective of serving as an annotated & curated database comprising complete genome sequence of viruses, value added derived data & data mining tools.

(2) The current release (v1.1) contains 559 complete genomes in addition to 287 putative genome of viruses belonging to eight viral families for which the host range included animals & plants.

(3) Viral genome in VirGen are annotated using sequence based Bioinformatics approach

(4) The genomic data is also curated to identify 'alternate names' of viral proteins where available

(5) VirGen archives the results of comparisons of genomes, proteomes & individual proteins within and between viral species

(6) It is the first resource to provide phylogenetic trees of viral species computed

using whole genome sequence data.

(7) The module of predicted B cell antigenic determinants in VirGen is an attempt to link the genome to its vaccinome.

(8) Comparative genome analysis data facilitate the study of genome organization & evolution of viruses which would have implications in applied research to identify candidates for the design of vaccines & antiviral drugs. VirGen is a relational database & ~~is available at http://bi~~

9) CORG database

(1) The Comparative Regulatory Genomics (CORG) database & annotation project aims at providing insights into gene regulation at the level of transcription.

(2) Having now several genomes of higher eukaryotes at hand, we are able to study sequence elements on a comparative basis.

(3) Comparative sequence analysis has become a powerful tool regarding a variety of problems ranging from gene finding to the identification of regulatory elements.

(3) The CORG project systematically applies methods to non coding genomic DNA

(4) The working hypothesis underlying the CORG project is the local sequence conservation points to functional importance.

(5) The CORG project is a resource for the genome wide annotation of conserved sequence elements in non coding genomic DNA. These elements are 'conserved non-coding blocks' (CNBs).

(6) ~~An~~ Sequence conservation upstream of two orthologous genes is likely to reflect similar regulatory control of the two genes and vice versa, many regulatory elements usually cluster in upstream regions

7) Thus, the CORG project focuses on detection & collection of conserved blocks from the upstream regions of orthologous genes.

8) We define non coding conserved blocks as local suboptimal alignments of non coding genomic DNA of orthologous

gene loci

9) To maximize the compatibility with ENSEMBL information on homology of gene loci was retrieved from a compilation of cross species gene relations in ENSEMBL.

10] The current CORE release 1.0 takes it non coding upstream sequences from NCBI Human Assembly 29 & the MUSC Mouse assembly 3.

a] Gene Prediction :—

① Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics

② Gene prediction basically means locating genes along a genome. Also called gene finding, it refers to the process of identifying the regions of genomic DNA that encode genes, RNA genes & other functional elements such as regulatory genes

Importance of Gene Prediction

(1) Helps to annotate large, contiguous sequence
(2) Aid in the identification of fundamental & essential elements of genome such as functional gene, intron, exon, splicing site, regulatory site, gene encoding known proteins, motifs, EST, ACR etc
(3) Predict complete exon - intron structure of protein coding regions.

Application

It has vast application in structural genomics, functional genomics, metabolomics, transcriptomics, proteomics, genome studies & other genetic related studies including genetic disorders detection, treatment & prevention.

First Generation Sequencing

OUTPUT

① The sequencing of clonal DNA populat

② low output

③ Short read sequencing

Accuracy

① Accuracy but limited scalability & high cost

② Accuracy > 99%.

Error

① Error rate of 0.001%.

Second Generation

OUTPUT

1] Reads around 36-600bp long

2] High Output

3] Short read sequencing

Accuracy.

1) High accuracy level

2) Accuracy > 99%.

Error

Error rate around 1%

3rd Generation Sequencing

OUTPUT:-

1) longer reads at lower costs & in shorter time
2) High output
3) Long read sequencing.

Accuracy

1) Accuracy 90-95%.

Error:

1) High error rate compared to other generations.