

Unit - 5

Genomics

Genomics

→ Gene - gene

→ omics - study in biology ending in the suffix - omic.

→ Genome - Hans Winkler (1920)
Coined the term

→ Sanger (1970s) - Genome - fundamental part of many omics
↑
Sequencing of genome.

→ word genomics is used in 1980's
widely used in 1990's.

→ Genome - genetic material of an organism.

↑
Coding genes

Non coding genes.

{ contains DNA genes
and Mitochondrial
& Chloroplast DNA.

→ Genomics - study of genome.

eg:- 46 chromosomes of humans
22 pairs - (Autosomes) - Somatic chromosomes
1 pair - Sex chromosomes (Allosomes)
46 chromosomes

↓
Study of genes in 46

Chromosomes - human genome.

- cells in a human body
↓
100 trillion
- chromosomes in a human cell
↓
46
- Genes in a human genome
↓
20,500
- DNA is - 2mks long.
- Genomics uses
- Recombinant DNA,
 - DNA sequencing method
 - Bioinformatics -
to sequence, assemble,
and analyze the structure
function of genome.
- ① human genome in humans
↓
99.9% same
- ② 1% of the genome contains genes
↓
2%
- ③ 1% of human proteins that share similarities with other organisms
↓
50%

Sequencing technologies:

- ① Sanger sequencing (1st generation sequencing)
- ② Next-generation sequencing (NGS)

① Sanger Sequencing:

- ① Fredrick Sanger with his colleagues developed (1977). - won Nobel prize in 1980.

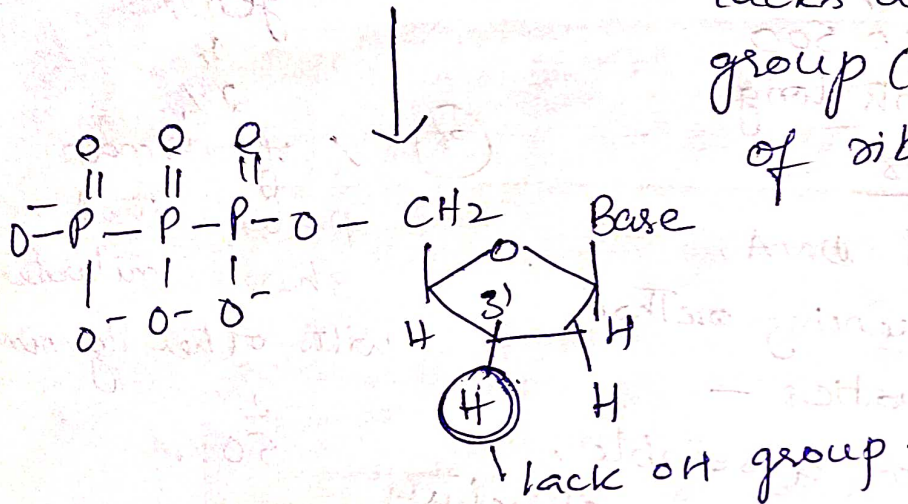
- Also called chain termination method,
(Dideoxy sequencing (DDNA Sequencing),
Enzymatic method)

- used in Human Genome Project
(2003)

→ These method is based on the chain termination by the use of ddNTPs ~~Di~~ dideoxynucleosides (ddNTPs)

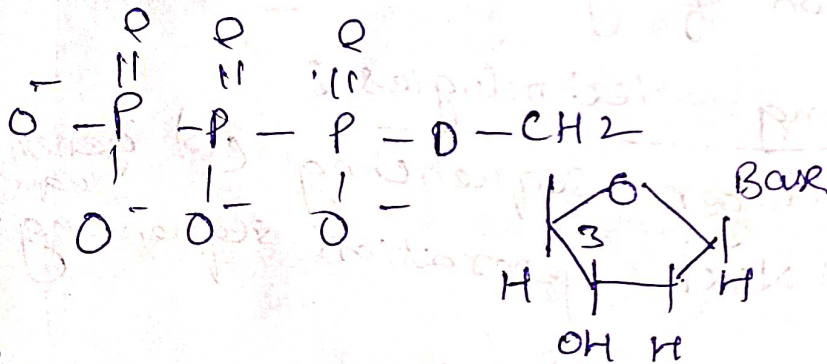
ddNTPs

→ A ddNTPs — Artificial molecule lacks a hydroxyl group (OH) at 3' carbon of ribose ~~sugar~~ sugar



↓
can't bind to nucleotide.

dNTPs



Requirements for Sanger sequencing

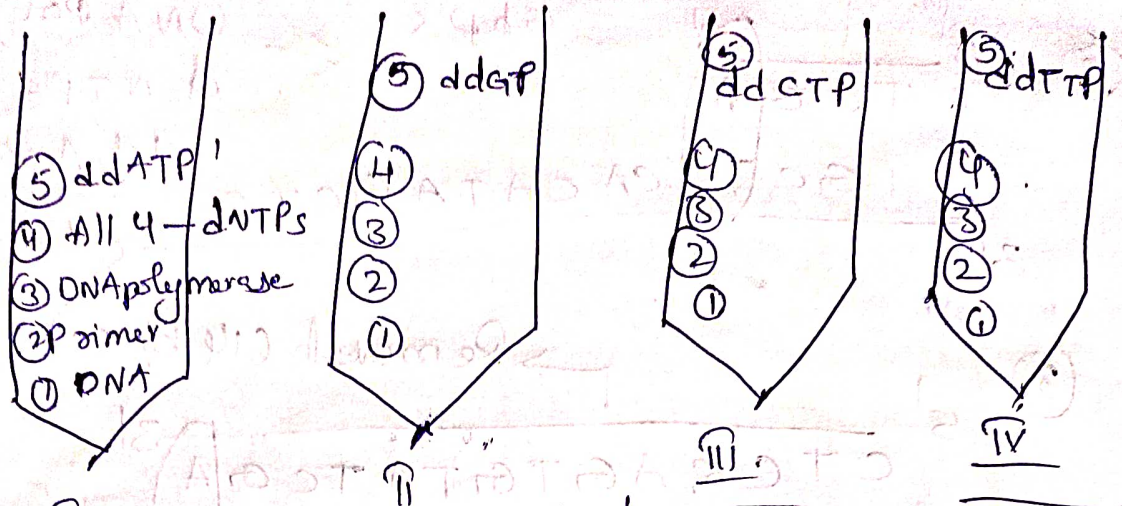
① DNA

② Primer

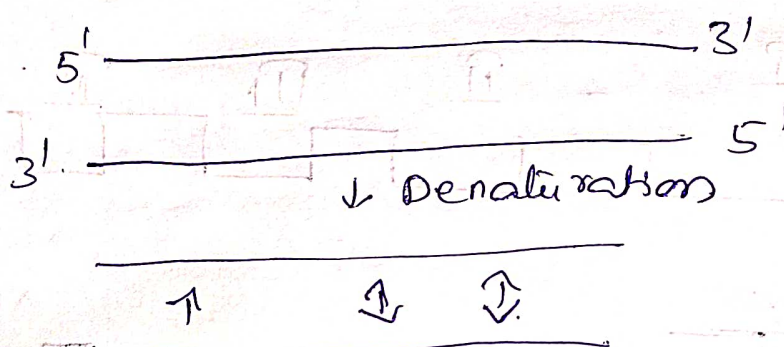
③ DNA polymerase

④ dNTP's [Deoxyribonucleoside triphosphate] dATP, dGTP, dCTP, dTTP

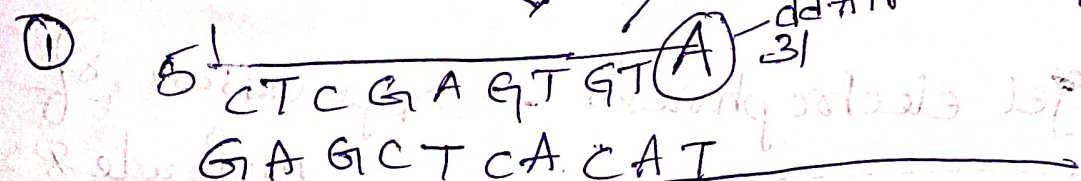
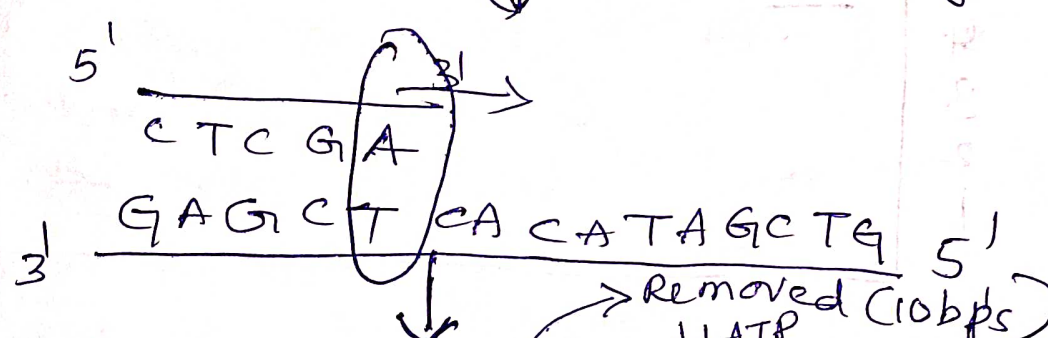
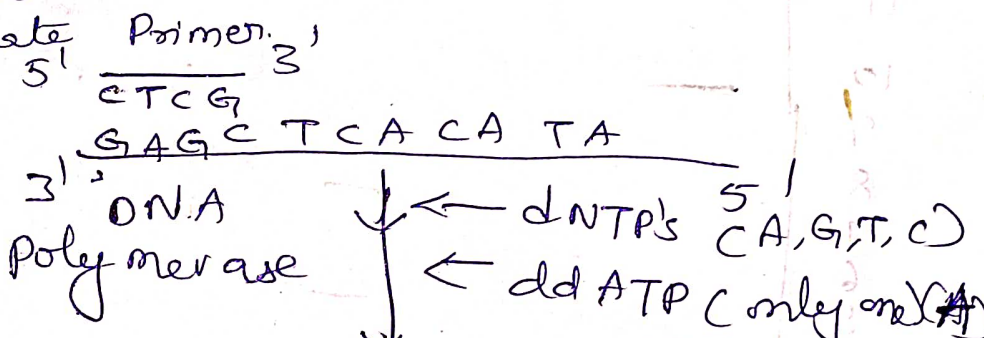
⑤ ddNTPs [Dideoxynucleoside triphosphate] ddATP, ddGTP, ddCTP, ddTTP

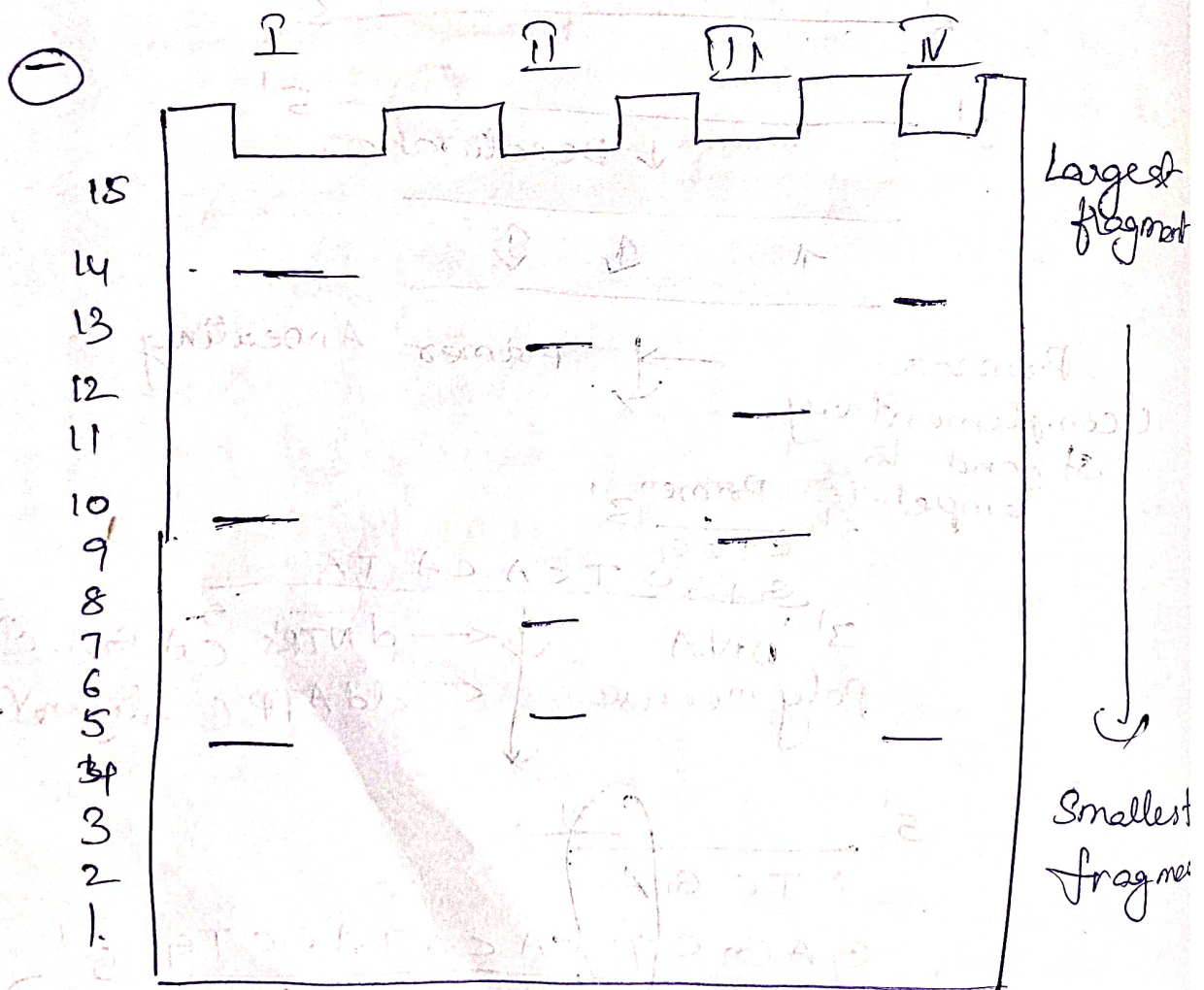
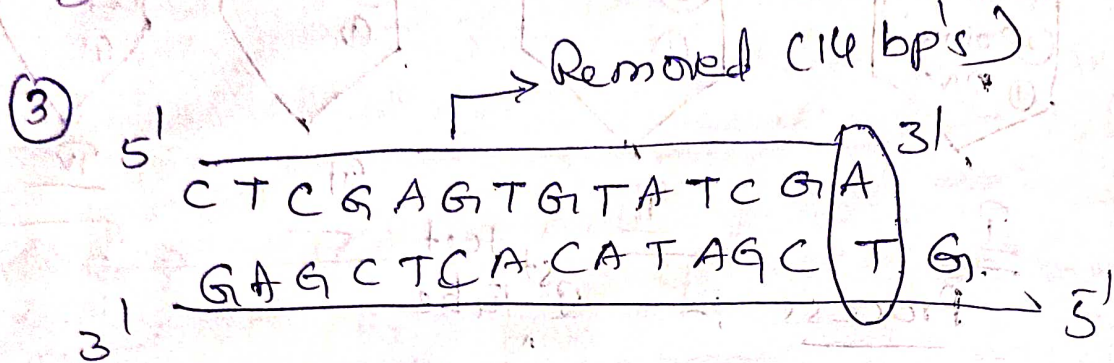
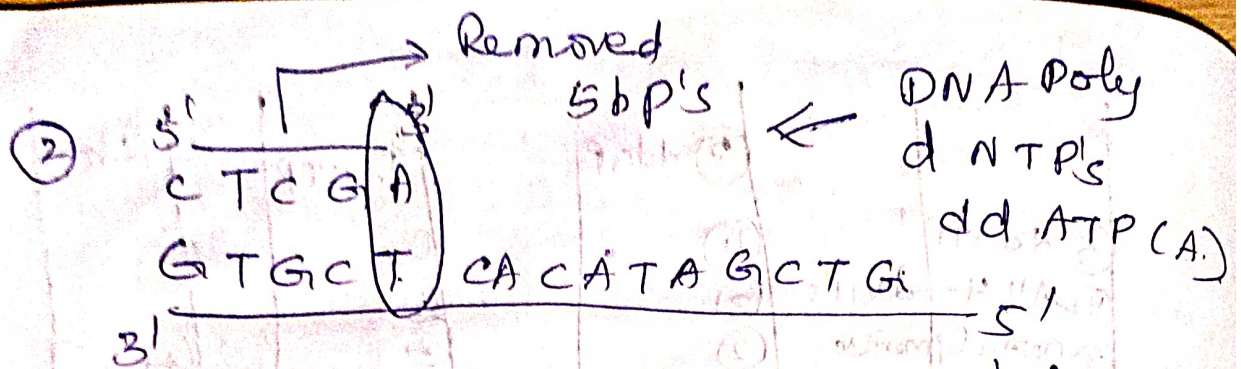


Process :- 1st test tube



Primer complementary strand to template → primer Annealing

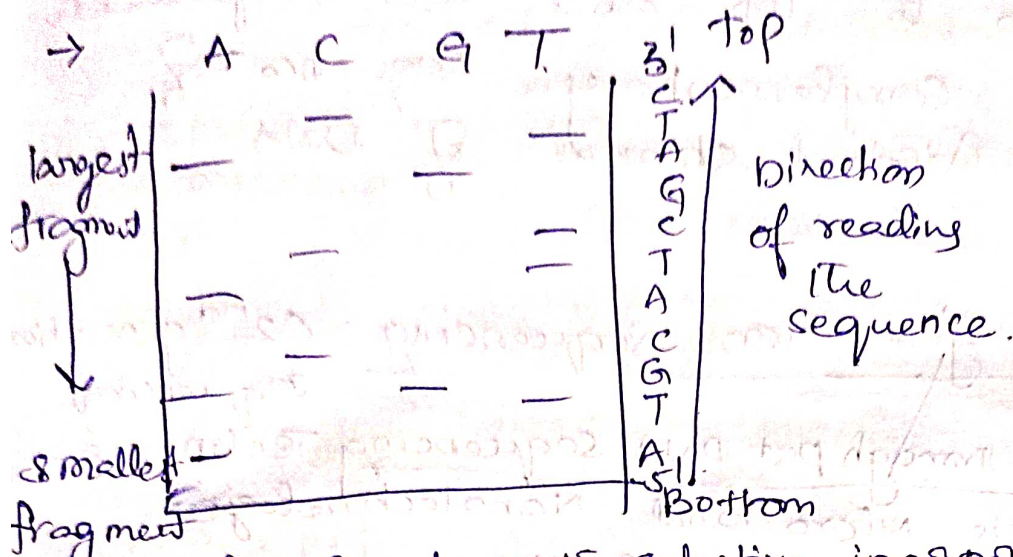




→ Gel electrophoresis — Separation of molecules

→ Agarose gel electrophoresis → Polyacrylamide gel electrophoresis

Autoradiography



Principle: Based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication.

Steps:

- ① A primer is annealed to a single-stranded section of DNA.
- ② DNA - primer mixture is put into 4 separate tubes with DNA polymerase and a solution of dNTPs at a concentration of 100 times lower than the dNTP concentration.
- ③ DNA pol uses dNTPs to extend the DNA.
- ④ ddNTPs are put together randomly, resulting in different lengths of fragments.
- ⑤ Fragments that are from each of the reactions are denatured and separated by size using gel electrophoresis.
- ⑥ The gel is used to visually detect the DNA fragments.

fragments are to be read from bottom to top and this represents the complementary sequence of the original strand of DNA.

② Next generation sequencing (2nd generation sequencing)

→ High throughput DNA sequencing technique

→ Employs Micro and Nanotechnologies:

a) Reduce Sample size.

b) Low Reagent cost

c) Less Time.

→ Massive parallel sequencing

→ Sequence thousands of sequences at once.

→ Produce enormous amount of data.

→ NGS Techniques.

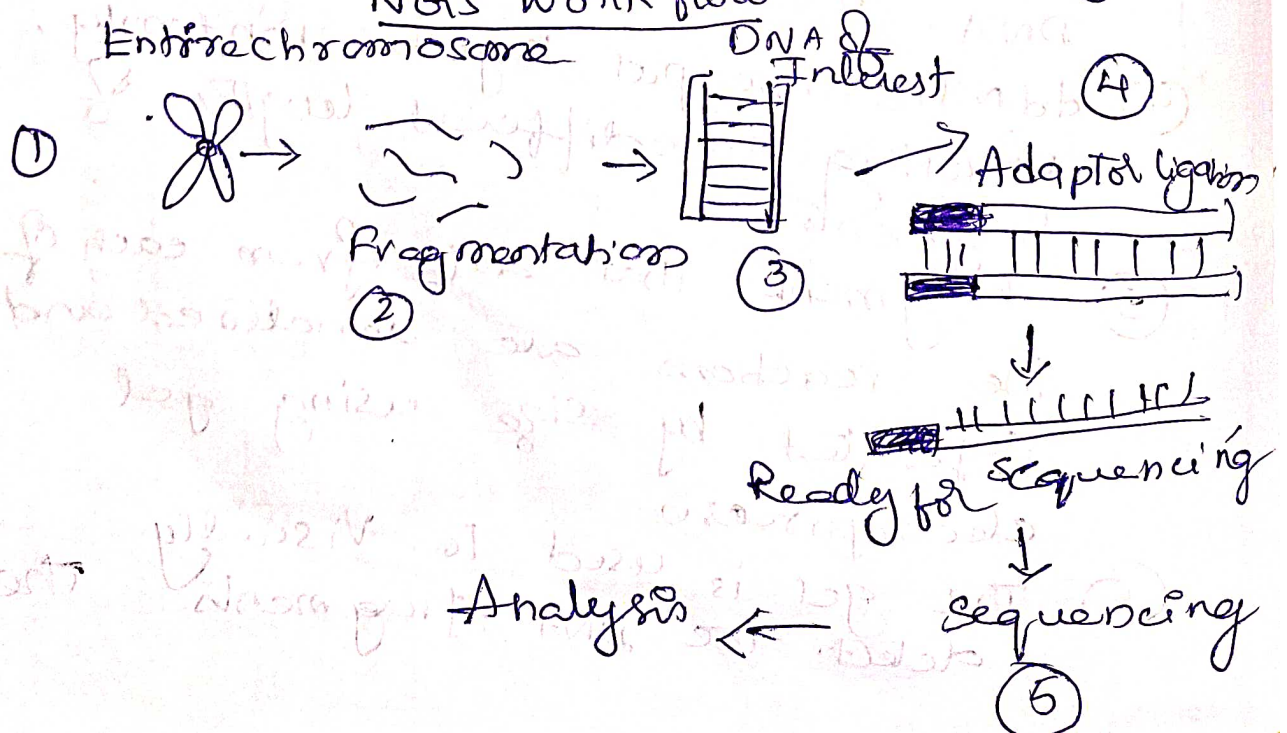
① Illumina sequencing

② Pyro-sequencing

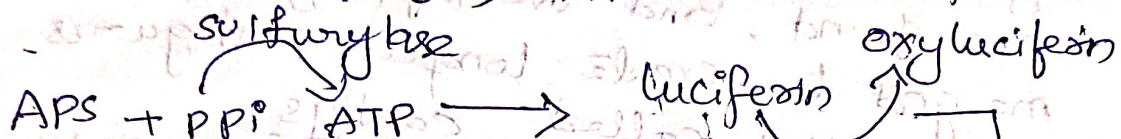
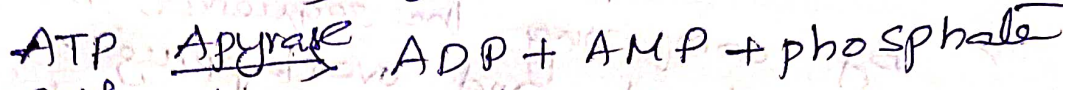
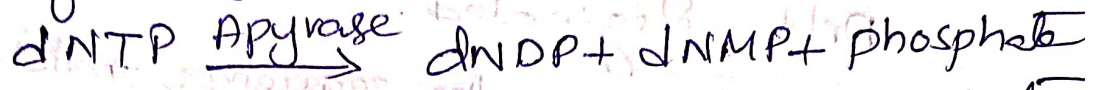
③ Roche 454 sequencing

④ Ion-torrent sequencing.

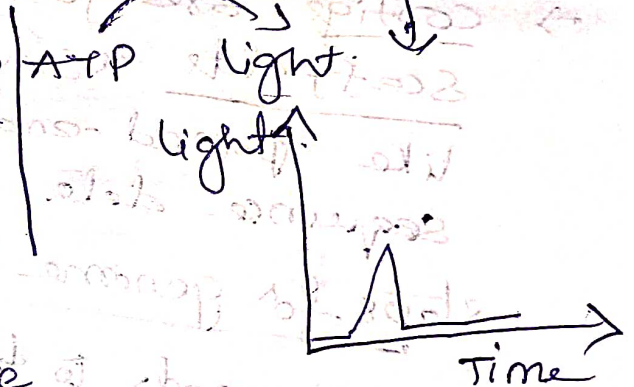
NGS work flow



① Pyrosequencing: Non-electrophoretic, bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light using a series of enzymatic reactions.



→ Nucleotide Incorporation generates light seen as a peak in the Program trace.



Key principles

- ① DNA sequencing: Libraries are generated by clonal amplification by PCR in vitro.
- ② DNA is sequenced by synthesis: Addition of Nucleotides
- ③ sequencing: Amplified DNA templates are sequenced simultaneously in a massively parallel fashion without the requirement for a physical separation step.

Steps involved in NGS

- ① Nucleic acid Extraction & Isolation
- ② Library Preparation
- ③ Clonal Amplification & sequencing
- ④ Data Analysis using Bioinformatics

Genome Assembly

→ The process of putting a large number of short DNA sequences back together to recreate the original chromosomes from which the DNA originated.

→ By comparing the sequences, matching regions are found, and overlapping reads are merged to create longer contiguous sequences called contigs.

→ Contigs are further assembled into larger scaffolds using additional information like paired-end reads or long-range sequence data.

Steps for genome assembly

- 1) Align reads to find overlapping regions
- 2) Determine a consensus sequence (or contig)
- 3) Scaffold contigs based on read pairs and/or overlapping regions
- 4) Generate pseudogenes based on genetic maps

DNA
↓
sequences
↓
contig
↓
scaffold.
↓
Mapped scaffolds.

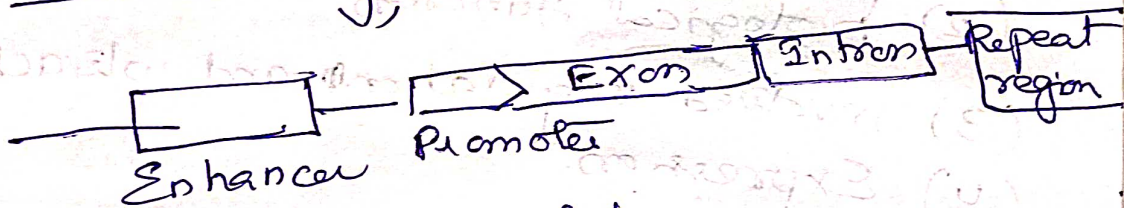
Genome annotation :- Dr. Owen White (1995)

→ It is the process of identifying the location of genes and discovering their function, as well as annotating other genetic features from a raw genome sequence.

→ Genome sequence

CATGACGTGCTACCA

↓
Genome Annotation



Genome Annotation

↓
Structural Annotation

→ Structure prediction

→ Identifying elements

→ Introns

→ exons, stop, start codons.

↓
Functional Annotation

→ function prediction

→ contains Biological information

↓
which protein is produced

Structural Annotation

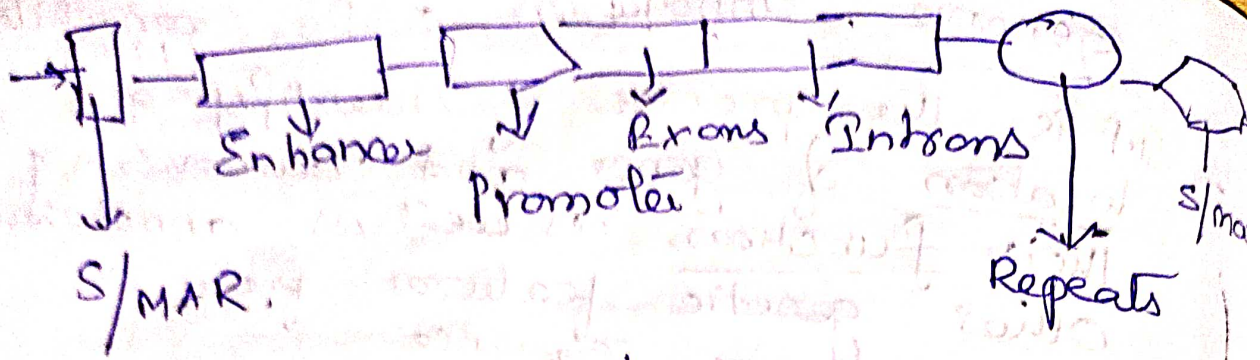
- identification of genomic elements

→ Open reading frame and their ~~localisation~~ localisation

→ gene structure

→ coding regions

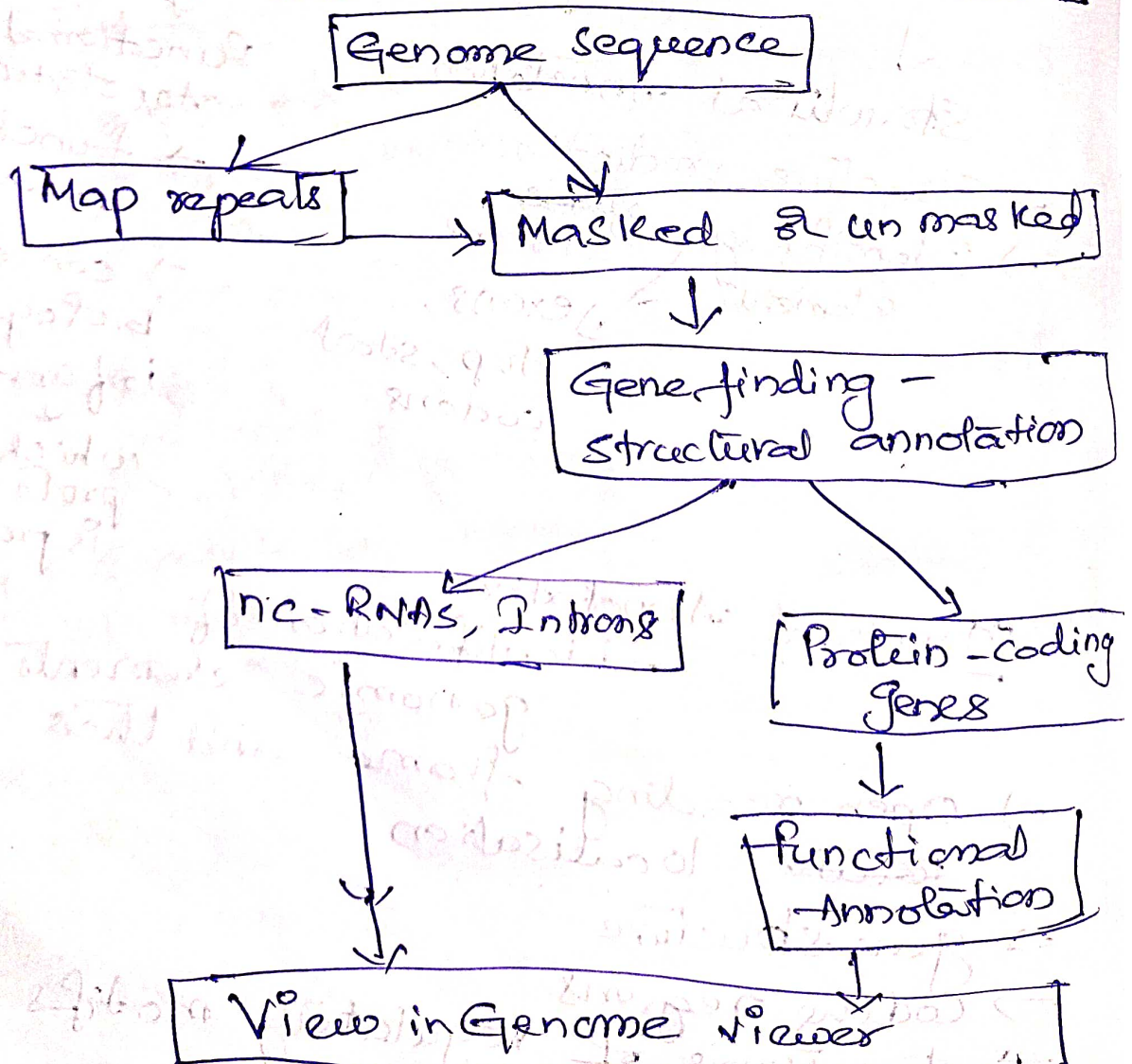
→ locations of regulatory motifs.



Functional annotation

- Attaching biological information to genomic elements.
- (1) biochemical function
 - (2) biological function
 - (3) involved regulations and interactions
 - (4) Expression.

Genome Annotation workflow



Genome of ~~Arabid~~ Arabidopsis (2M)

- Arabidopsis — 2000
Genome project (1990-2000)
 - ↳ Cruciferae
 - ↳ Rapid growth
 - ↳ low DNA content
 - ↳ small genome size
- contains about 125 megabases of sequence
- Encodes approximately 25,500 genes
- Contains a similar number of gene functional classifications as other sequenced eukaryotic genomes (Drosophila melanogaster & Caenorhabditis elegans)
- has 85% unique genes
- has 39.5% genes that exist as members of large gene families
- shows evidence of ancient polyploidy as estimated 58-60% of the Arabidopsis genome exists as large segmental duplications.
- Analysis of sequence of the Arabidopsis genome → Similar to multicellular organisms.
- It also differs from animals
- Plants are autotrophic → Require light, H_2O , minerals
 - ↓
 - photosynthesis.
- But animal genome differ from plant genome due to ~~plant~~ Heterotrophic mode of nutrition.
- Arabidopsis centromeric regions — heterochromatic
 - ↓
 - 4.7 expressed genes are prot.

- Arabidopsis has unique genes. — approximately 150 unique protein families.
- Arabidopsis genome contains genes encoding RNA polymerase subunits not seen in other eukaryotic gene. 16 unique families of transcription factors.
- Arabidopsis — contains several classes of proteins — used in animal systems.

~~Arabidopsis~~

Difference b/w Sanger sequencing & NGS

Sanger Sequencing

- ① Sequence Volume. → ① low sequence volume
- ② Sensitivity → ② low
- ③ Speed & Cost. → ③ Slow turnaround time with high cost per base.
- ④ Time taken for completion → ④ long time
- ⑤ Sequencing → clones, PCR samples

NGS Sequencing

- ① millions of reactions in parallel, resulting in high-throughput
- ② High.
- ③ Quick ~~turnaround~~ time with low cost per base.
- ④ Short time
- ⑤ DNA libraries

- ⑥ preparation steps → ⑥ few, sequencing reaction clean up Many, Complex procedures
- ⑦ Data collection → ⑦ samples in plates: 96, 384 samples on slides 1-167

Applications of NGS

- Bioinformatic analysis
- NGS → hundreds of megabases to gigabases of nucleotide sequence at a time
- genome sequencing — used in biomedical sciences
- low cost of sequencing

Genome of Mycobacteria

- Mycobacteria — vet gram bacteria
- Mycobacteria — 1882 — Robert Koch
- Robert Koch — discovered along with M. Tuberculosis, cholera vibrio, Bacillus anthracis.
- won nobel prize — 1905

characters

- gram +ve bacteria, 4.4 million base pairs.
- slow growth
- Dormancy
- Complex cell envelope
- Intracellular pathogenesis
- Genetic homogeneity

species

M. africanum — Humans
M. bovis — cattle.
M. microti — voles
M. canettii — Humans
M. tuberculosis — Humans.

Aim :-

- Transmission, clinical ~~phen~~ phenotypes
- develop advanced vaccine against the TB
- develop drug

steps :- Involved in Genome sequencing
Genomic DNA



Cosmid / BAC library



Shotgun sequencing



Genome sequence



Relational Database

→ cosmid — (1.4 - 2.0 kb)

BACs — were cloned into M13 Vectors.

→ less than 0.1% — clone coverage

< 0.2%, from one strand

with sequence chemistry

(Informatics

→ data base — EMBL, TREMBL, SWISS PROT
BLASTN, BLASTX, FASTA

→ M. tuberculosis — 4,411,529 bp

— G+C content — 65.6%
— Genome Size 4.4 Mb (4,411,529 bp)