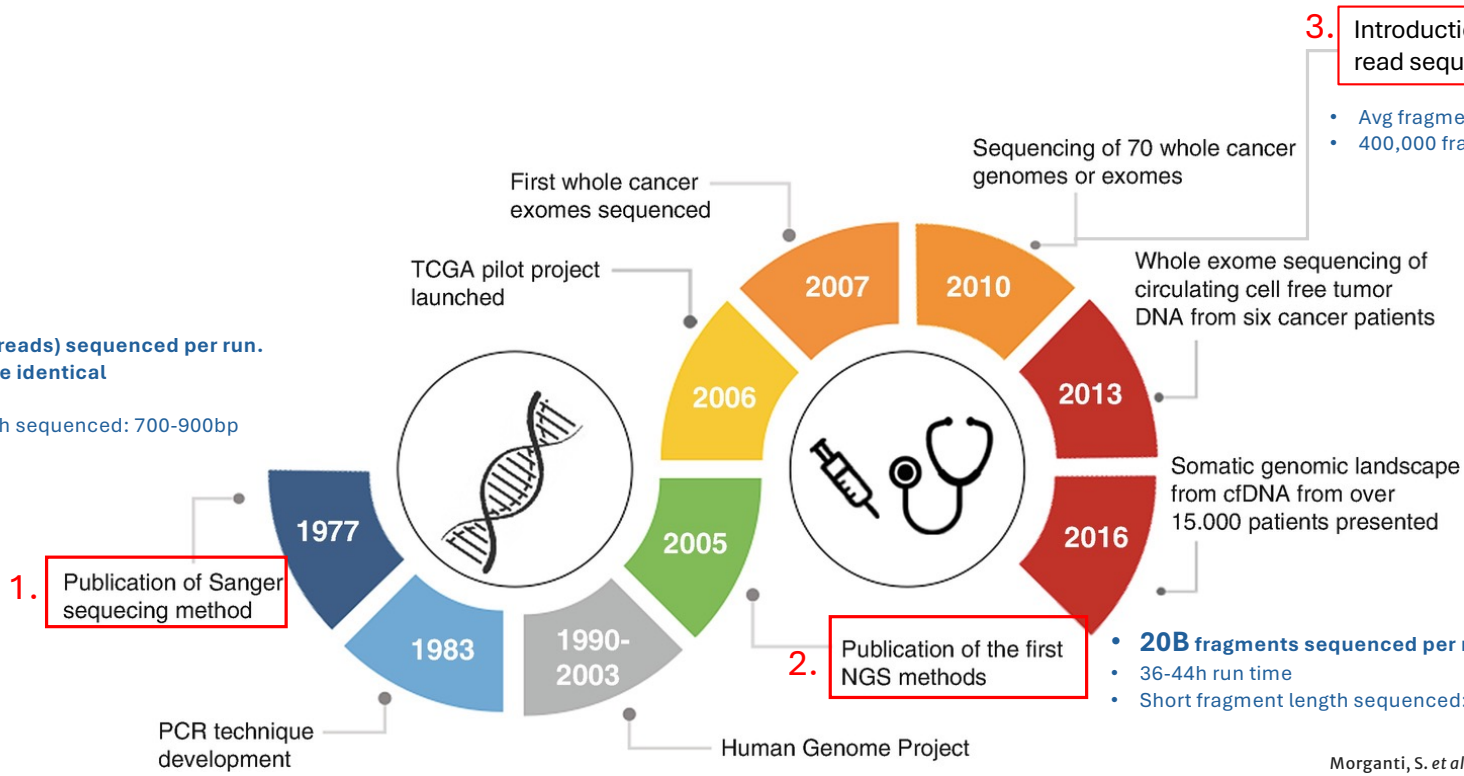
THEORETICAL BASIS FOR SEQUENCING

Evolution of sequencing technology and its influence in Scientific Discovery and Medicine

- **96 DNA fragments (reads) sequenced per run.**
- **DNA fragments were identical**
- 2-3h run time
- Short fragment length sequenced: 700-900bp



- Avg fragment length sequenced: **30,000 bp**
- 400,000 fragments sequenced per run

- **20B fragments sequenced per run (10K fold more!)**
- 36-44h run time
- Short fragment length sequenced: 300-500 bp

Morganti, S. *et al.* (2020). Role of Next-Generation Sequencing Technologies in Personalized Medicine. In: Pravettoni, G., Triberti, S. (eds) P5 eHealth: An Agenda for the Health Technologies of the Future. Springer, Cham. https://doi.org/10.1007/978-3-030-27994-3_8

Outline

First generation | **Second generation** (next generation sequencing) | **Third generation**

Sanger sequencing
Maxam and Gilbert
Sanger chain termination

Infer nucleotide identity using dNTPs, then visualize with electrophoresis

500–1,000 bp fragments

454, Solexa,
Ion Torrent,
Illumina

High throughput from the parallelization of sequencing reactions

~50–500 bp fragments

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

Tens of kb fragments, on average

PacBio: Mar25/2020: Sequencing 101
https://www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

Short-read sequencing

Past but certainly **not obsolete**

- Fast, cost -effective and highly accurate for **low throughput sequencing**
- Longer read length than NGS and simpler than long read sequencing
- Mainly used to confirm:
  - Genes being cloned
  - PCR products
  - Site directed mutagenesis
  - DNA fingerprinting for cell line authentication
- **No longer meets sequencing demand for high-throughput sequencing** nor is it cost-effective

Present the **bread and butter** of sequencing

- The work horse of sequencers
- Used for WGS, transcriptomics (RNAseq), epigenetics, single cell RNAseq, spatial RNAseq and even some spatial proteomic studies
- Disadvantage: short read sequencing

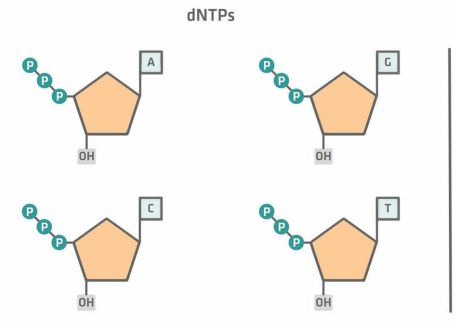
Long-read sequencing

Future?

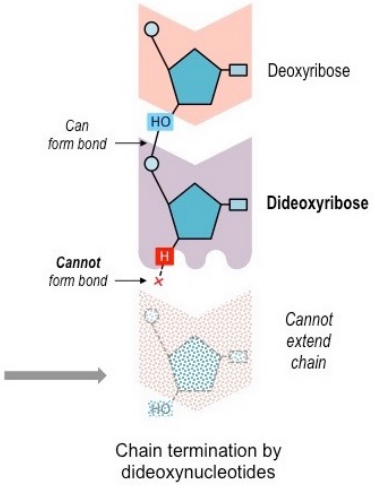
- Advantages: long reads can improve de novo assembly, mapping certainty, transcript isoform identification, and detection of structural variants
- Disadvantages: high error rate in sequencing, cost and longer time required for data analysis

Sanger Sequencing- 1st generation sequencing

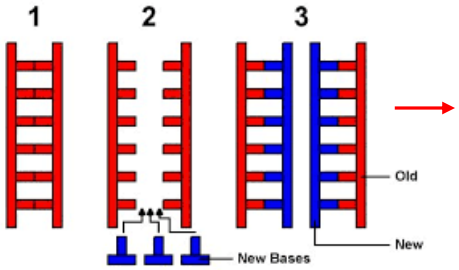
- Chain termination sequencing which in part relies on **PCR methods** and a special type of nucleotide call **dideoxy NTP (ddNTP)**



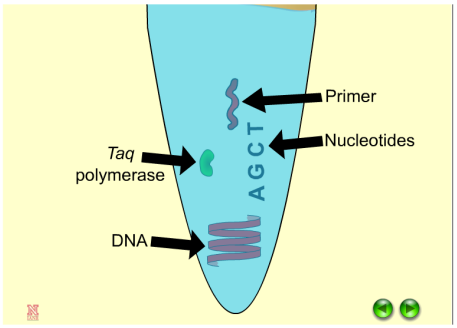
PCR2: random incorporation of ddNTP into growing nucleotide chain resulting in termination of reaction

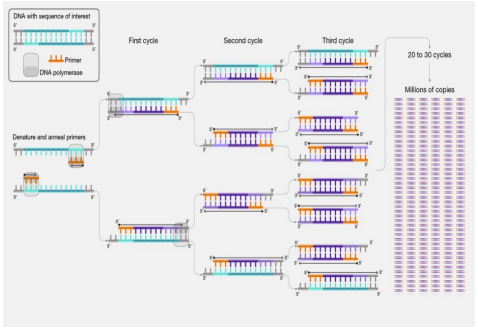


Replication in cells:

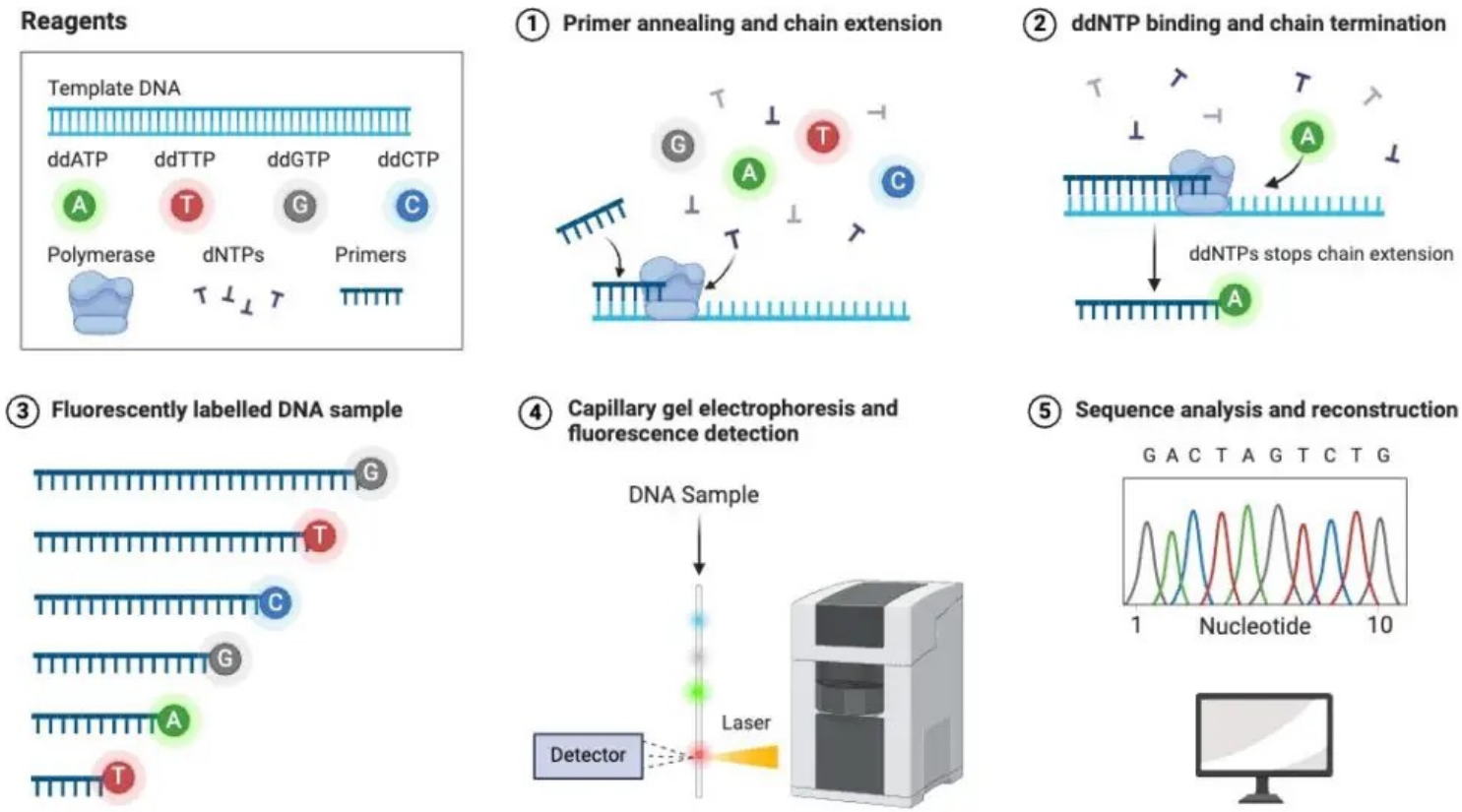


PCR1: DNA fragment amplification





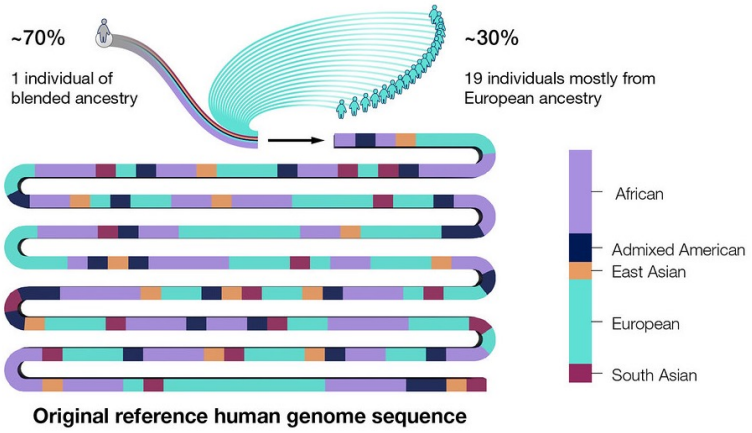
Sanger Sequencing- 1st generation sequencing



https://www.yout-ube.com/watch?v=dVRB4CaLizc

https://www.youtube.com/watch?v=X9566yl2cBo

Human Genome Project



WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

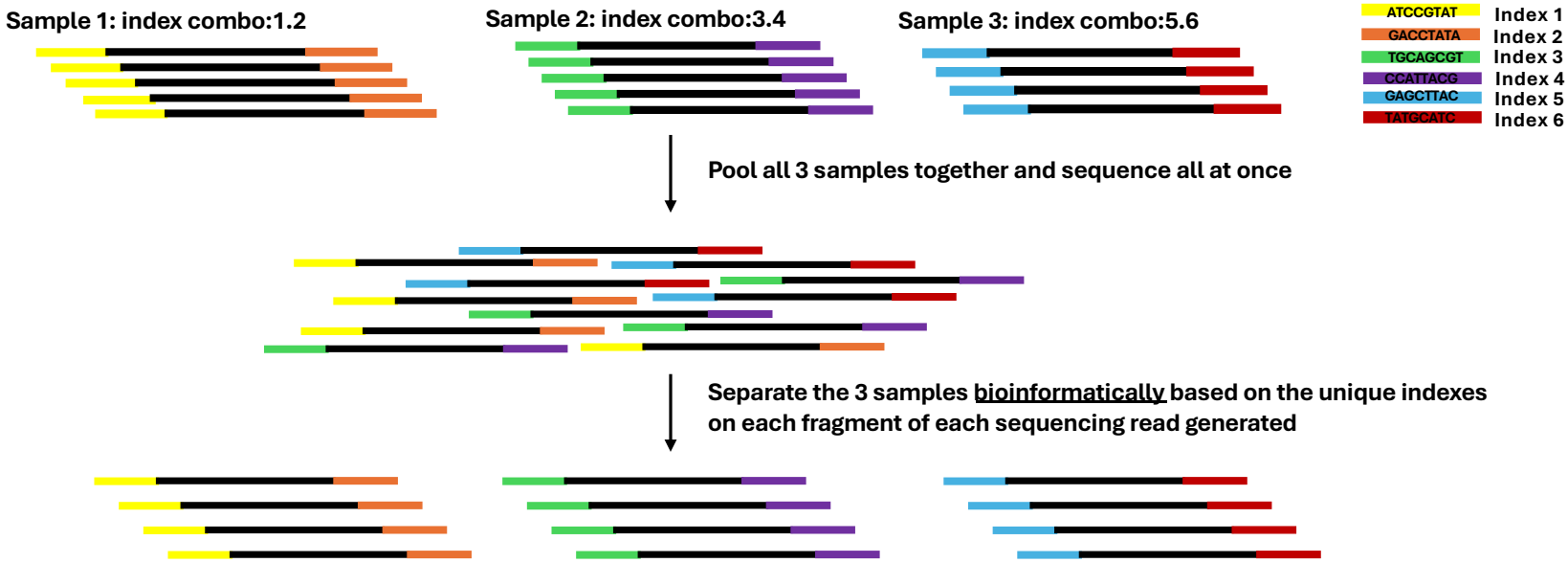
ROSWELL PARK CANCER INSTITUTE

For more information please contact the Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

- The Genome Project was completed 2 yrs ahead of schedule due the dedication of an international consortium of researchers from US, UK, Germany, France, Japan, China
- the de facto leader was the US scientist Dr. Collins from the National Human Genome Research Institute (part of NIH)
- 92% of the genome was sequenced at the time, the rest, around 400 gaps, were difficult to sequence due the limitation of technology at the time
- **Gaps were highly repeated regions or sequence, often found around the region of telemores and centromers**
- The enter genome was not actually completely sequenced until 2022 largely due to the advent of long red sequencing
- The ENTIRE genome was finished by T2T consortium co-lead by Dr. Karen Niga (UCSC) and Adam Phillippy, a bioinformatician, from NHGRI

Next Generation Sequencing (NGS)- 2nd generation sequencing

- Simultaneous, massively parallel sequencing where 100s of 1000s of DNA fragments are **being sequenced at the same time** instead of one at a time like Sanger sequencing
- Not only can an ENTIRE **genome** or **exome** or **transcriptome** be sequenced at once **but many individual samples of each can be combined** and sequenced in one run.
- This is done by **indexing** all the DNA fragments in one sample with a short unique barcode of 8-12 nucleotide bases



- **A huge amount of data is generated in a short period of time and the method of parallel sequencing is cost effective and well as accurate**

Next Generation Sequencing (NGS)- 2nd generation sequencing

Timeline of technological advancements

2005: Roche-pyrosequencing (SBS)

2007: Helicos-Heliscope

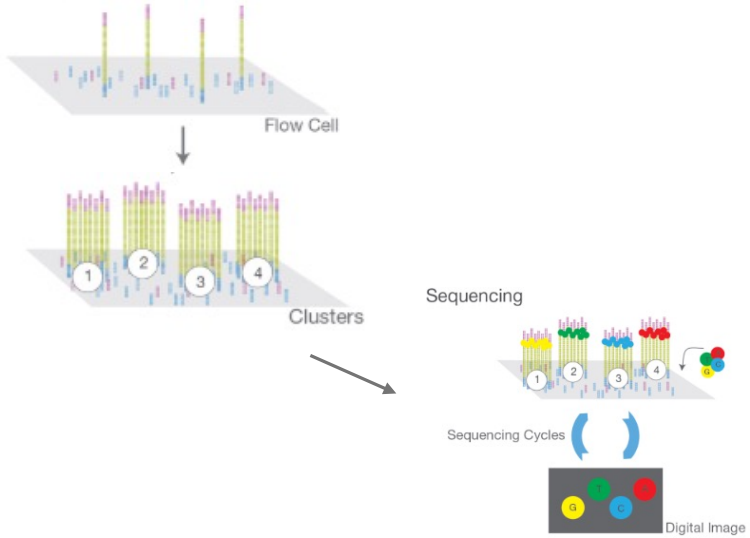
2006: Illumina-Genome analyzer(SBS)

2007: ABI SOLiD

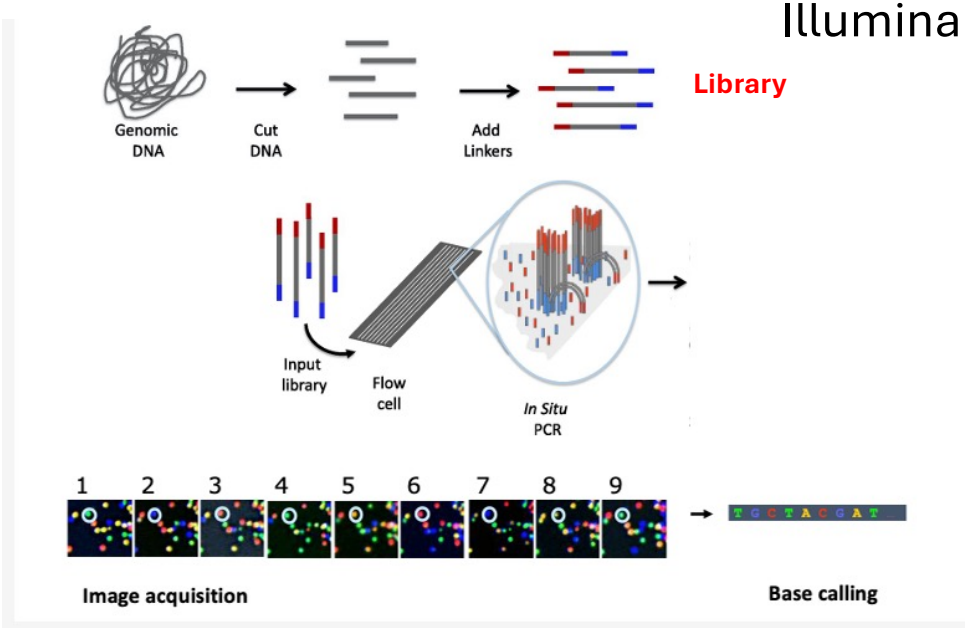
2010: Ion Torrent

Ion Torrent: https://youtu.be/zBPKj0mMcDg

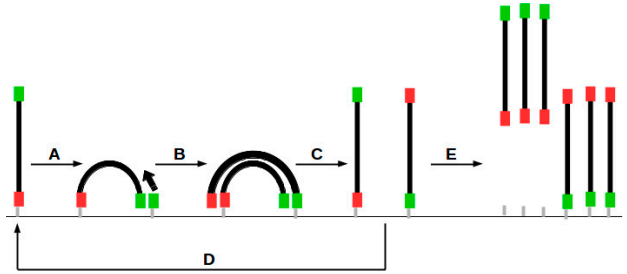
- SBS- sequencing by synthesis
  - DNA fragments bind to flow cell and PCR1
  - PCR based **SBS- sequencing by synthesis**
- Why did Illumina become the gold standard for NGS sequencing?
  - Cluster amplification technology enhanced the accuracy of base calling which helped reduce the need for sequencing redundancy and reduce thereby cost
  - Better optics generated stronger signals



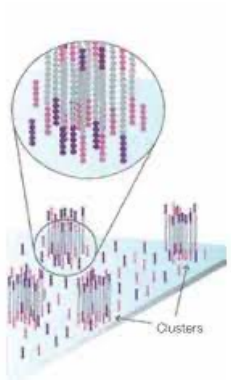
Illumina Sequencing

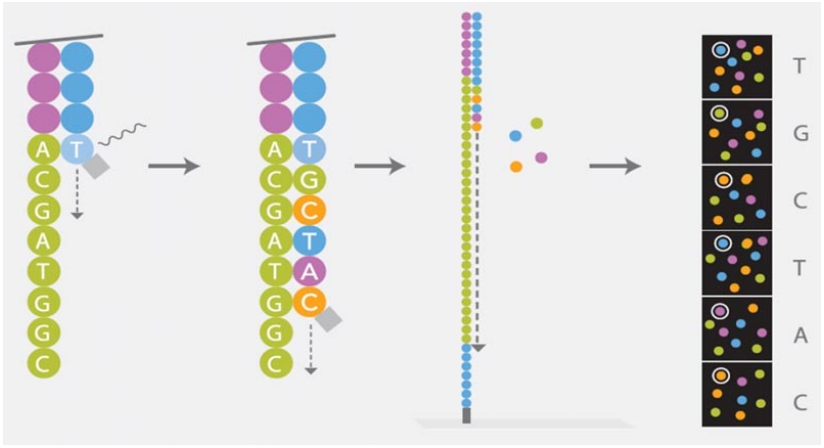


Library



Cluster generation





Sequencing:

1. Primer binds
2. Flow cell flooded with fluorescently- tagged dNTPs w. blocker
3. Only **ONE** dNTP binds per DNA fragment (bc of blocker)
4. Wash away excess dNTP
5. Records florescence
6. Cleave blocker = completion of cycle 1
7. Repeat (usually 50-300 cycles)

https://www.youtube.com/watch?v=fCd6B5HRaZ8

https://www.youtube.com/watch?v=EDVKxSNdSic

Library Preparation- prepare DNA for sequencing

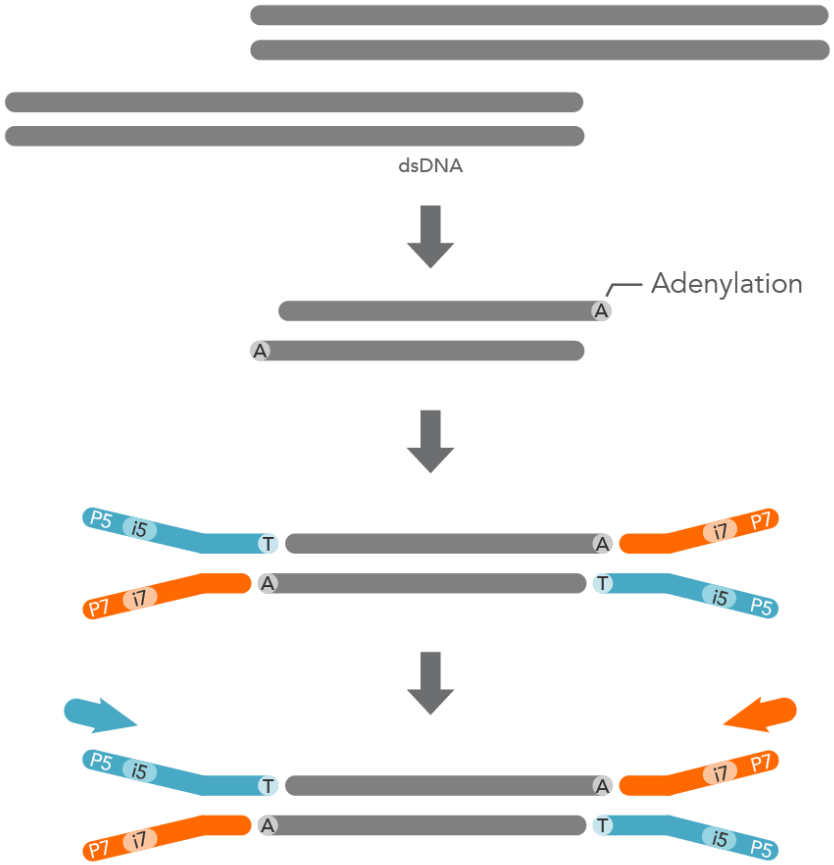
1. DNAseq: WGS, WES, amplicons

Fragmentation

End repair and A-tailing

Ligation

PCR amplification



Fragmentation:

- Enzyme
- Mechanical
- transposase

Adenylation:

- it activates the DNA ends
- provides the necessary energy
- ensures the reaction occurs with high specificity.

Adapters:

- Flow cell binding site (linkers)
- Sequencing primer binding site
- Unique indexes to identify individual samples (added during PCR step)

DNAseq

- Many types of DNAseq:
  - WGS- whole genome sequencing-
    - Determine the complete DNA sequence of an organism's genome
    - Provides comprehensive view of the organism's genetic make-up
    - Includes both introns (non-coding) and exons (coding)
    - detects a broad spectrum of genetic variations, including single nucleotide variants, insertions, deletions, and structural variants, which may be missed by targeted sequencing methods.
    - Disadvantage: cost, complex data analysis, ethical concerns
  - Exome sequencing-
    - Exome sequencing focuses on the exons-only the coding regions, rather than the entire genome, making it a more efficient and cost-effective approach.
    - WGS library 1st made, exomes are enriched and then sequenced
    - Disadvantage: can miss variants that may be in the non-coding regions that could influence regulation of gene expression
  - Amplicon sequencing- 1st PCR amplify targeted regions of the genome which are of interest which are then sequenced

Types of Library Preparations

2. RNAseq- Transcriptomic studies

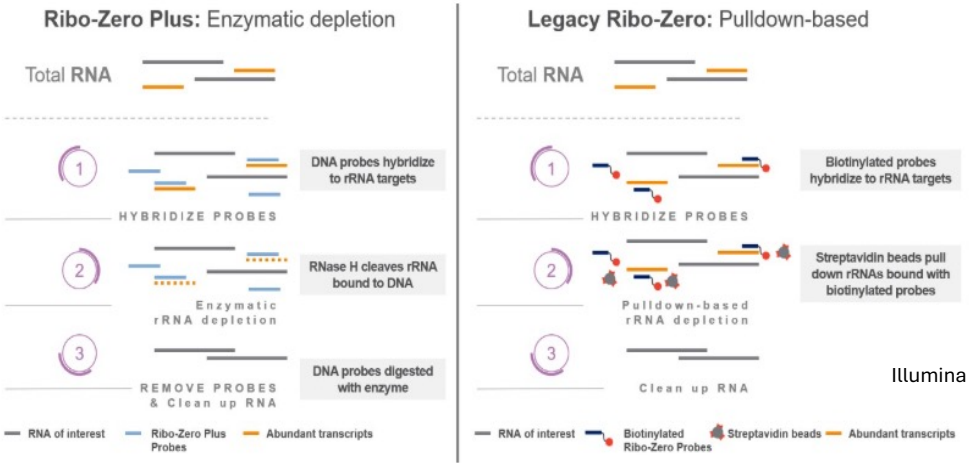
2 PROBLEMS:

- Most of our RNA in cells is rRNA- conserved, won't see many differences- want mRNA and must remove rRNA
- RNA is unstable- can't sequence directly

POSITIVE SELECTION OF mRNA: mRNAseq (capture mRNA using polyA tail and KEEP)

- Reverse transcriptase- generates ssDNA also known as 1st strand synthesis
- DNA polymerase- generates double stranded DNA = cDNA

NEGATIVE SELECTION OF mRNA: TotalRNAseq (capture rRNA using probes and DISCARD)



Illumina

- General library preparation workflow (previous slide)

IDT Technologies

RNAseq and the study of Transcriptomics

- Transcriptomics- study of the whole transcriptome (all the genes that are actively made into mRNA) as it changes over a variety of biological changes

- Driving force of how disease state or abnormal biological state FUNCTIONALLY differs from normal state in Genomics

- Many types of RNAseq:
  - mRNAseq- positive selection via polyA tail
  - total RNAseq- negative selection by eliminating the rRNA using RNAseH or biotin-streptadivin-bead system
  - miRNAseq- (regulation)
  - RIP-seq- RNA IP (regulation)
  - single cell RNAseq
  - spatial RNAseq

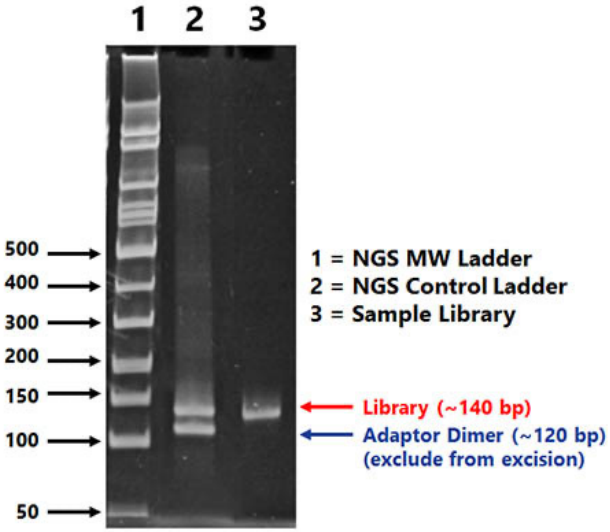
miRNAseq-microRNA

miRNA Function in post-transcriptional regulation of mRNA (inhibit or degrade)

Isolate TOTAL RNA

Proceed with library preparation as normal

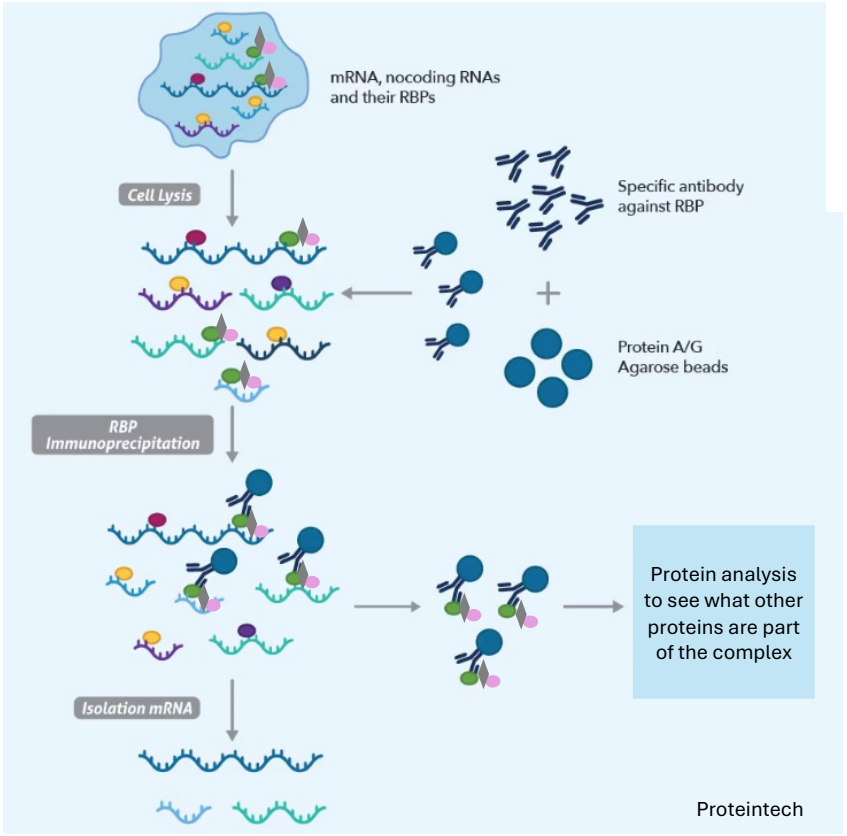
After PCR, perform gel electrophoresis to excise bank corresponding to miRNA

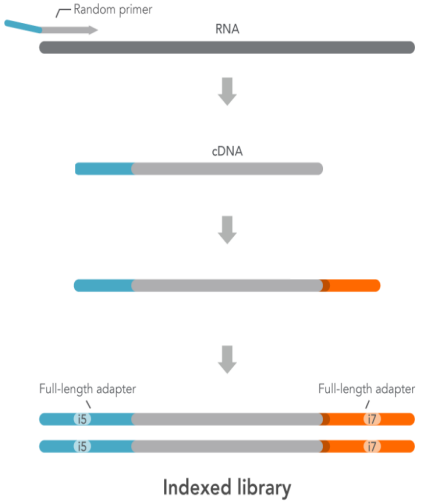


Purification of the DNA in the band which corresponds to miRNA

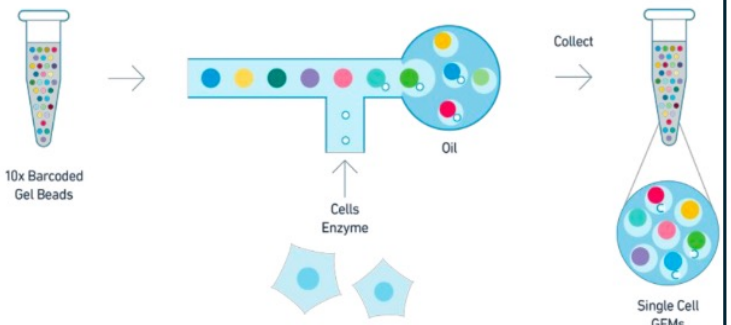
RIPseq

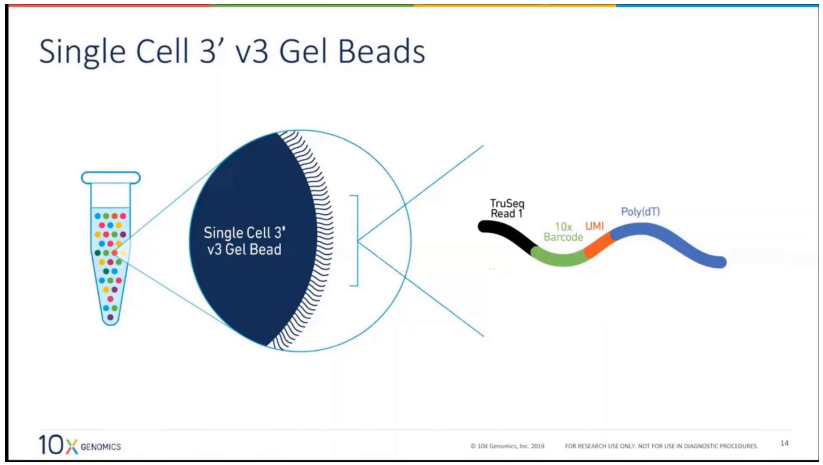
Maps RNA binding sites for proteins.
Crucial for various cellular functions, including DNA repair, RNA splicing, protein synthesis, and gene regulation, and disruptions in these interactions can lead to diseases.





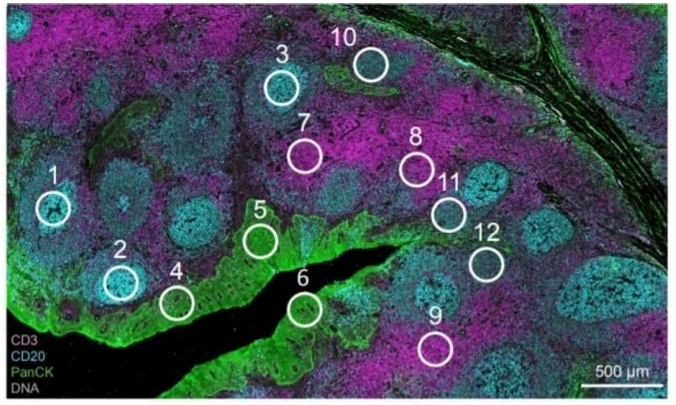
Single-cell RNAseq (Droplet seq)

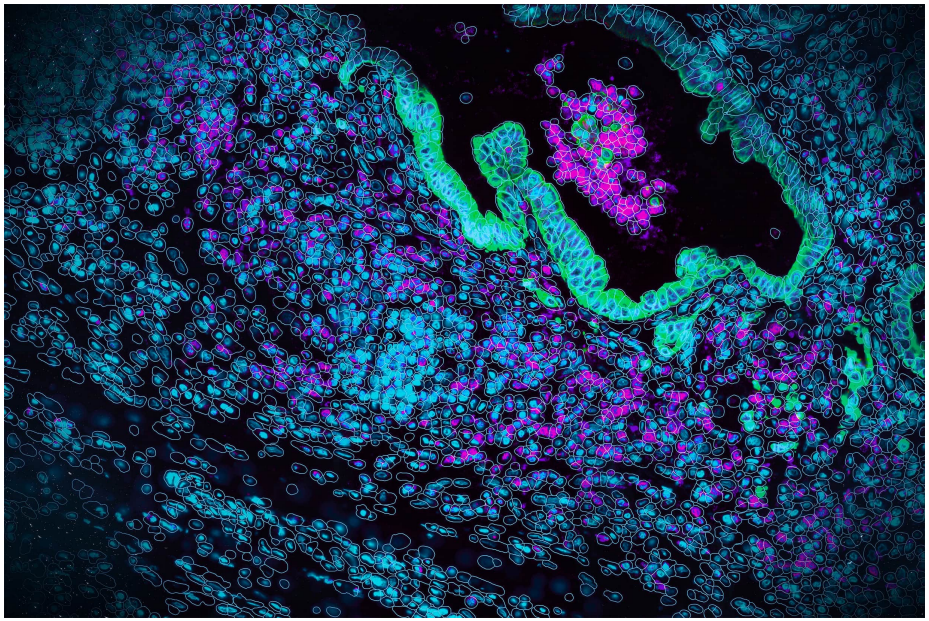




10x Genomics

Spatial sequencing





Nanostring (formerly)
Bruker Corporation

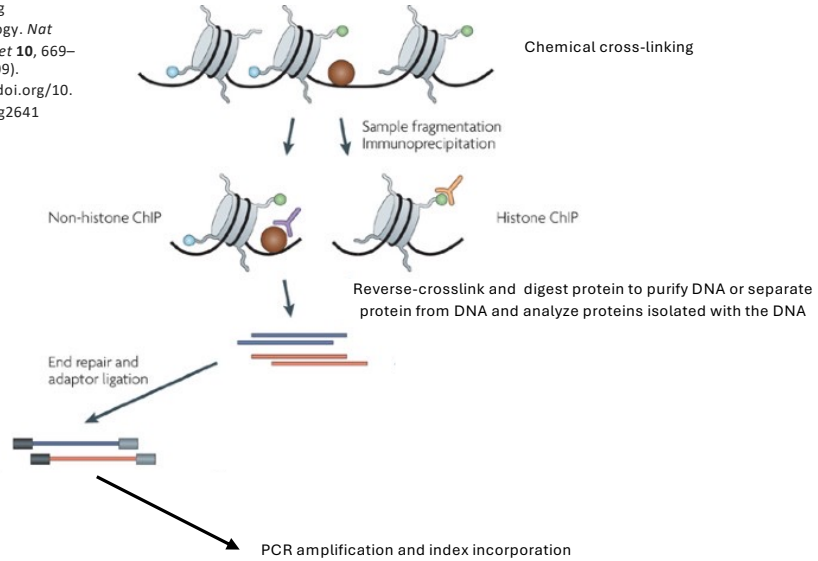
Types of Library Preparations

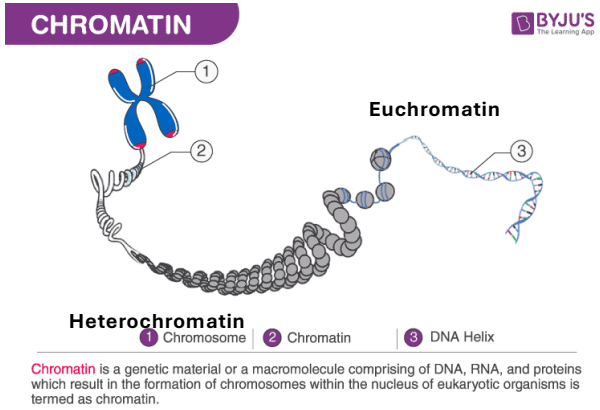
3. DNAseq-epigenetics

- Epigenetics- study of chromatin structure, organization and function, primarily in gene regulation
- ChIPseq- chromatin IP. Offer insight into protein-DNA interactions. Histones
- ATACseq- maps chromatin accessibility
- Hi-C- investigates 3D chromatin organization

ChIPseq- (chromatin IP):

Park, P. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669–680 (2009). https://doi.org/10.1038/nrg2641





2 types of Chromatin in cells:

1. Euchromatin:
   - **Definition:** Euchromatin is the less condensed, more open form of chromatin.
   - **Function:** Euchromatin is actively transcribed, meaning the DNA within is being used to make RNA and proteins.
   - **Modifications:** Euchromatin is often associated with histone acetylation, which helps to loosen the structure and allow access for transcription machinery.

2. Heterochromatin:
   - **Definition:** Heterochromatin is the more condensed, tightly packed form of chromatin.
   - **Appearance:** It appears as dark, dense regions in the nucleus.
   - **Function:** Heterochromatin is generally inactive, meaning the DNA within is not being transcribed.

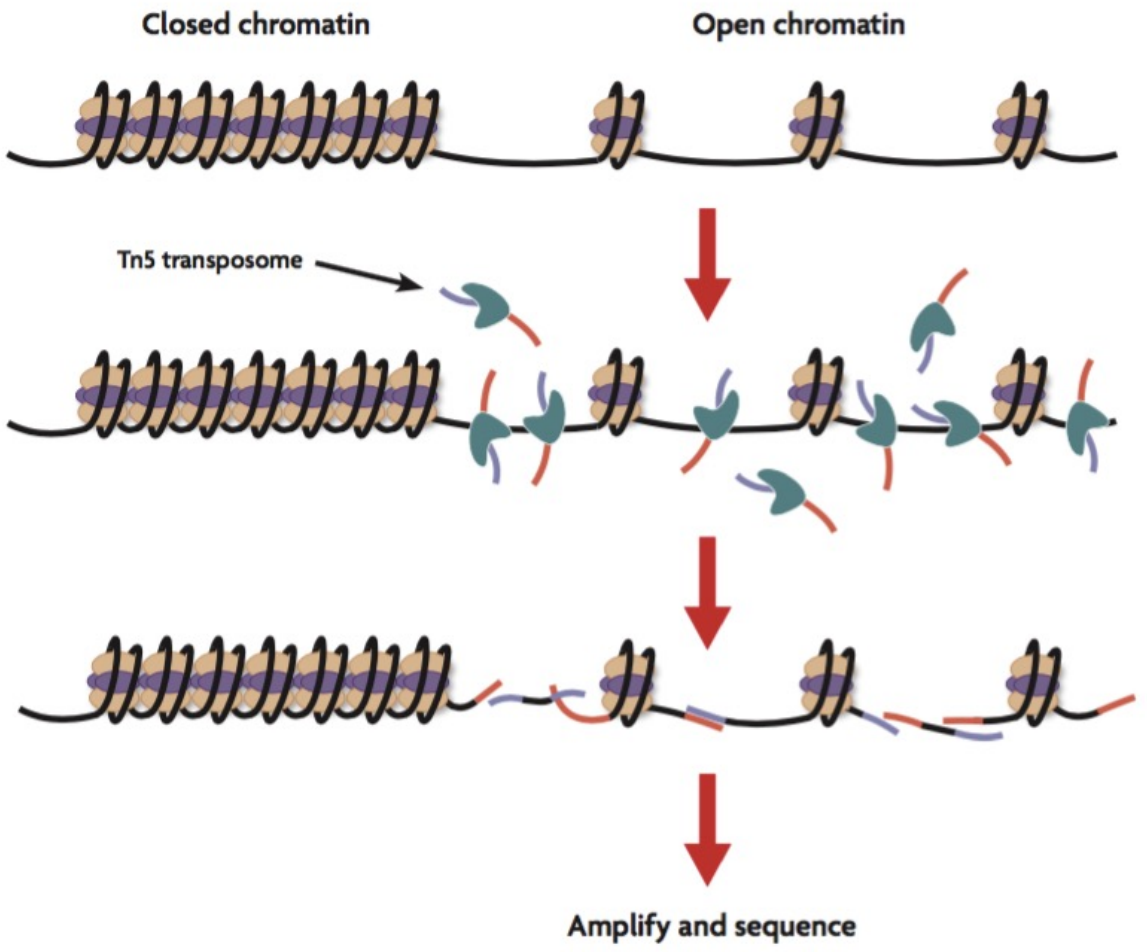
Histones modification can give us insight on what type of genes are associating with the histone.

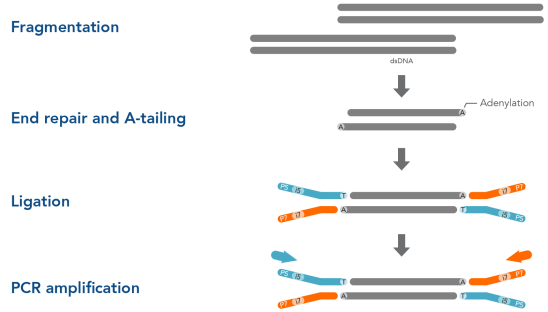
Ex. H2K9Ac histones are often associated with genes that function as gene activators

Chromatin immunoprecipitation that will pull down genes associating with H2K9Ac may identify genes involved in activation in a sample

https://youtu.be/992RkrUwGfo?si=g24nko9bNq71Z_Oi&t=68

ATAC seq- (Assay for Transpose Accessible Chromatin):





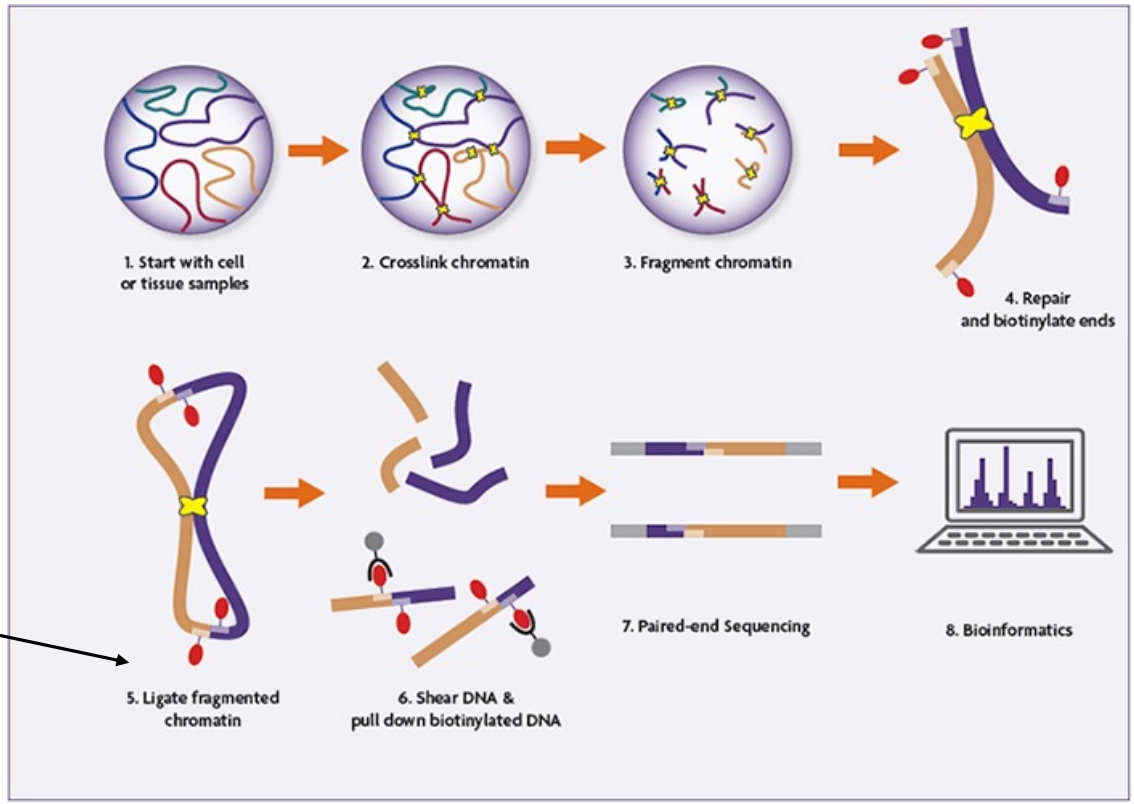
Active Motif

Hi-C- (Assay for Transpose Accessible Chromatin):

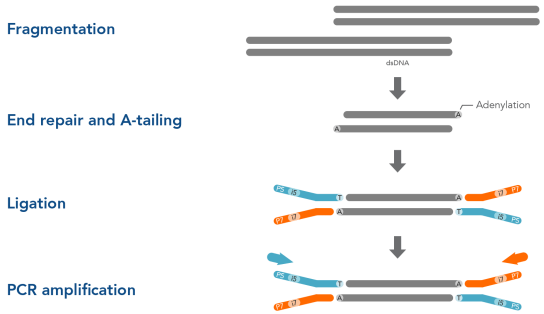
a method to study the three-dimensional structure of genomes by mapping chromatin interactions genome-wide, revealing how distant genomic regions interact in the nucleus

Measure how 2 fragments of DNA may interact with each other

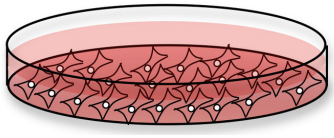
Studying the 3D structure of genomes and its relationship to gene regulation and evolution.



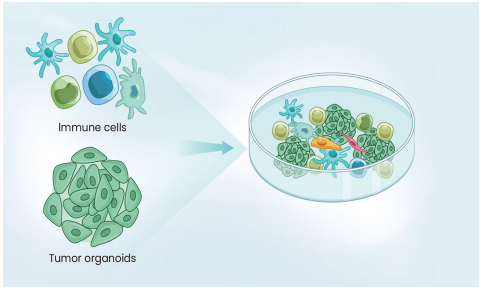
Proximity ligation



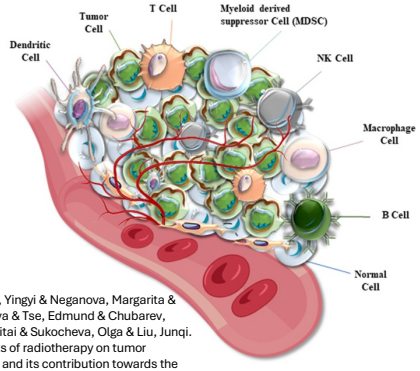
VERSALITLITY OF SEQUENCING



Cells in culture

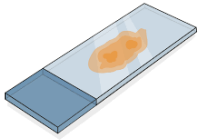


Cells in a 3D co-culture (organoids or spheroids)
Mimic the 3D environment of cells in our bodies



Zhao, Deyao & Mo, Yingyi & Neganova, Margarita & Yu.R., Aleksandrova & Tse, Edmund & Chubarev, Vladimir & Fan, Ruitai & Sukocheva, Olga & Liu, Junqi. (2023). Dual effects of radiotherapy on tumor microenvironment and its contribution towards the development of resistance to immunotherapy in gastrointestinal and thoracic cancers. Frontiers in Cell and Developmental Biology. 11. 1266537. 10.3389/fcell.2023.1266537

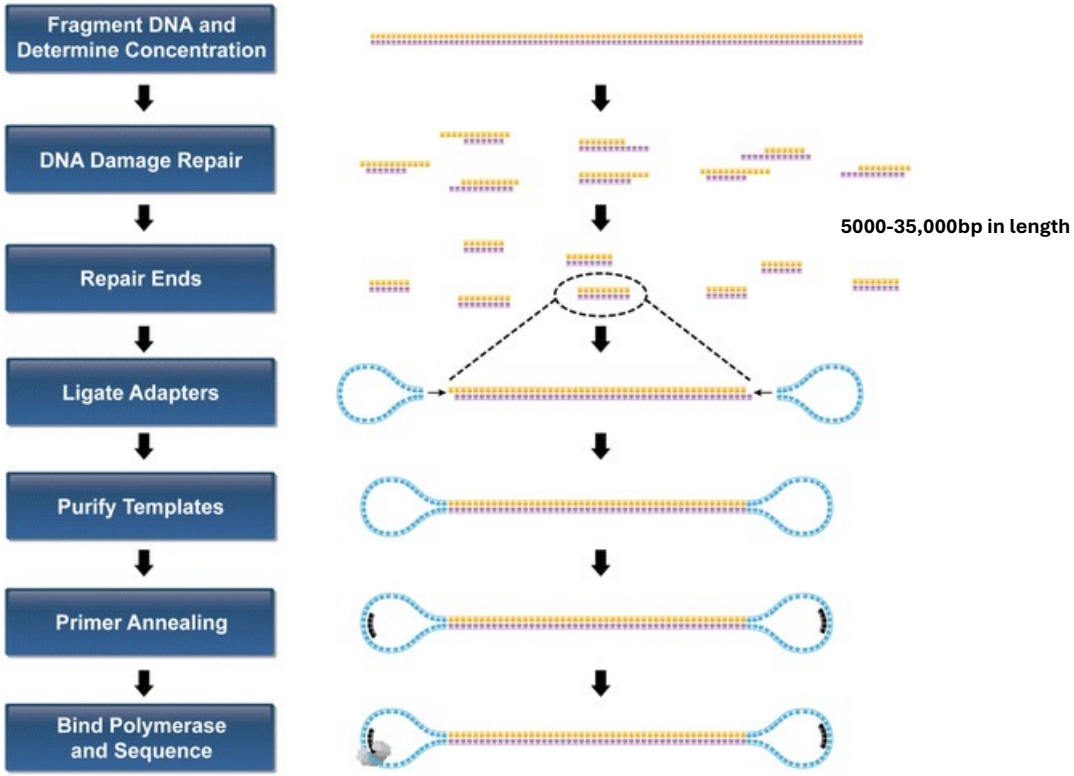
DNA/RNA from blood, serum/plasma, biopsies

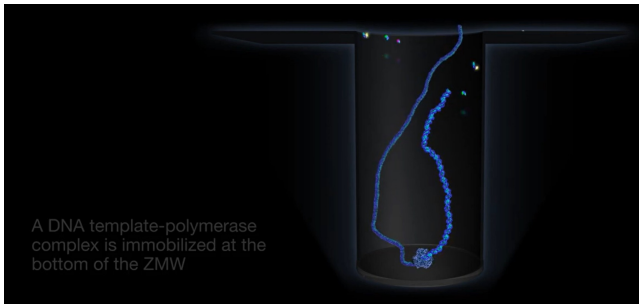


Long Read sequencing- 3rd generation sequencing

- Avg read length 1000-30,000 bp (unlike 2nd generation 300-500bp)
- Longer read length minimizes gaps in sequencing which is common using NGS in highly repetitive regions and areas of structural variation
- Single and native (no PCR amplification) DNA/RNA molecule being sequenced
- First technology came out in 2010-11. **PacBio** SMRT technology (Single Molecule sequencing in Real Time)
- Between 2011 and 2018 PacBio increased output of 1000 bp per read to 30,000 bp per read!
- In 2014 **Oxford Nanopore** released its first portable nanopore sequencing device, the MinION.
- The overall data output was small but enough to **identify pathogens**
- This made it possible for DNA sequencing to be carried out almost anywhere, even with limited resources.
- **A quarter of the world's SARS-CoV-2 viral genomes were sequenced with nanopore devices.**
- Both are fast but error rate higher than NGS, this has gotten much better in the recent years and there is lots of excitement about long read sequencing
- Cost is still an issue, especially with PacBio. ONT is making headway in this aspect making long read sequencing more affordable.
- With ease in cost, researchers can overcome higher error rates by doing redundant sequencing.

PacBio- SMRT





ZMW (zero-mode waveguide) is a nanophotonic device, a tiny well or hole in a conductive layer, that confines light and allows for the real-time detection of nucleotide incorporation events during DNA sequencing.

https://youtu.be/_lD8JyAbwEo

https://youtu.be/NHCJ8PtYCFc

PacBio

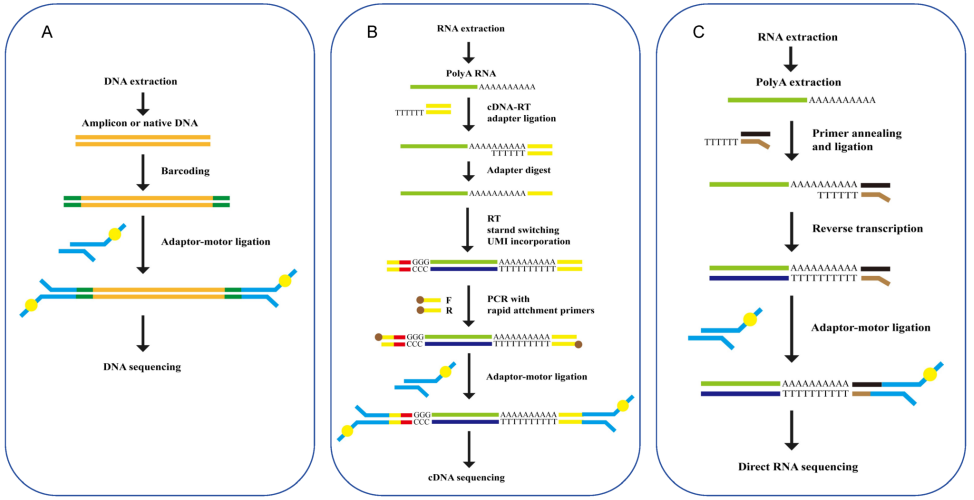
Fonte: Kong N, Ng W, Thao K, Agulto R, Weis A, Kim KS, Korlach J, Hickey L, Kelly L, Lappin S, Weimer BC. Automation of PacBio SMRTbell NGS library preparation for bacterial genome sequencing. Stand Genomic Sci. 2017 Mar 23;12:27. doi: 10.1186/s40793-017-0239-1, PMID: 28344744; PMCID: PMC5363030.

Nanopore- ONT

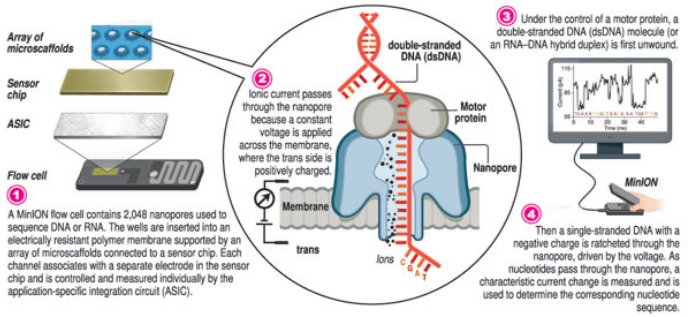
- Unlike all previous sequencing technologies, ONT doesn't use any DNA polymerases
- Instead, it uses a barrel-shaped protein called α-hemolysin – naturally found as a 'pore' in a cell membrane, this is called a nanopore
- α-hemolysin has a diameter of 1 nanometre – just big enough to allow a single strand of DNA through
- α-hemolysin nanopores are embedded into an artificial membrane inside a flow cell
- When a current is applied to the membrane, the DNA travels through the nanopore
- As the DNA travels through the nanopore, it obstructs the current flowing across the membrane and makes a 'squiggle'
- The four bases of the DNA (A, T, C and G) are of different shape and size, so each base has its own variation of the 'squiggle'
- These variations are measured by an electronic chip. An algorithm converts the data into a sequence which can then be read.
- Unique feature of nanopore: **adaptive sampling**

https://www.youtube.com/watch?v=RcP85JHLmnI

https://www.youtube.com/watch?v=fwHreHs9FHg



Viruses **2024**, *16*(5), 798; https://doi.org/10.3390/v16050798



Nanopore sequencing – rapid insights in real time

By Kumudini Hettiarachchi and Ruqyyaha Deane
View(s): Sunday Times Jan 2022

SUMMARY

- 1st generation- Sanger. Sequencing 1 DNA fragment at a time. Read length 700bp
- 2nd generation- Illumina. Massively parallel sequencing of 100 or 1000s of DNA fragments at one time. Read length: 300-500bp
- 3rd generation-PacBio or ONT- long read sequencing. Read length 1000-30,000 bp
- Majority of sequencing in the last 2 decades have been 2nd generation sequencing, with Illumina monopolizing the market
- WGS, Chromatin accessibility and interaction, mRNA expression in bulk, single cell and spatial platforms have given enormous insight into disease states, cancer and has as revalorized Genomics