15/01/2024

Next-Generation Sequencing

- → Roche 454 (2006-2013)
- → Illumina → most popular (90%)
- → PacBio SMRT → single-molecule real-time
- → Ion Torrent
- → Oxford Nanopore → latest technology

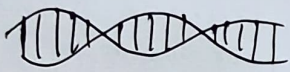
→ Data Analysis
→ Genomics
→ Transcriptomics
→ Assembly Algorithms

Sanger sequencing method:

di-deoxy termination method.

↓

capillary electrophoresis



- ✓ one DNA fragment at a time → Drawbacks
- ✓ we get around 1 kb (10^3 bp)
- ✓ took 13 years to sequence the entire human genome!

Human Genome Project

Shotgun sequencing

NGS Technologies:

- ✓ direct sequencing
- ✓ high-throughput, accurate ~ 96 fragments together
- ✓ cost-effective ~ $1000 & can go down

Atlas Project

↳ expression of genes changes during diseases → The Cancer Genome Atlas (TCGA). ← dataset to correlate genetics & drugs

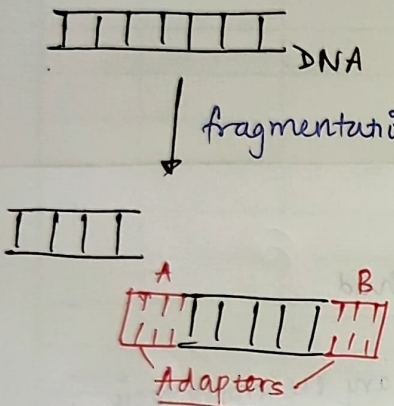
Precision Medicine ← useful in patient-specific therapy.

Reads:

- ✓ output sequences from sequencing platforms
- ✓ of different lengths

Roche 454 Sequencing:

↳ Pyrosequencing



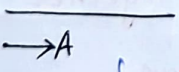
↳ synthetically-designed DNA sequences of known sequences → primer-binding sites → as we do not know the seq. of fragment yet

↳ Index sequence/ DNA barcode → to identify each sample differently

Multiplexing → mixing multiple samples & sequencing them together

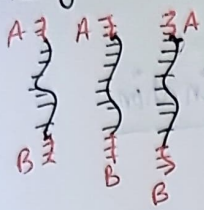
Pyrosequencing:

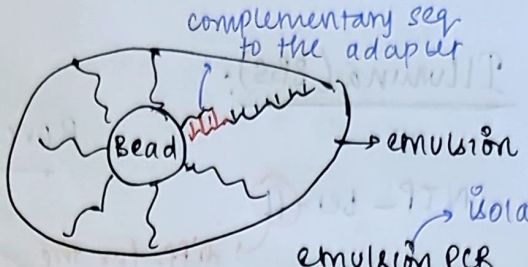
↳ relies on synthesizing the complementary strand



↳ every time there is an addition of a base, a light signal is emitted → if two same bases simultaneously → double the intensity

fragment



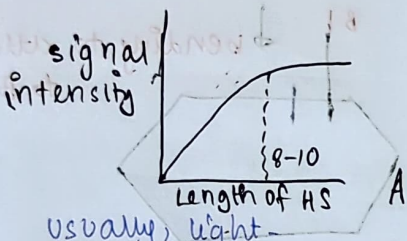


emulsion PCR → isolation of fragments → clonally amplified.

picotitre plates → 100,000 fragments → put in sequencer

Drawbacks:

- ✓ many enzymes & substrates required
- ✓ expensive compared to Illumina, etc.



ATTTTTTTT...

homopolymeric stretches ~5-20

usually, light intensity increases proportionally with repeated bases

after ~8 repeats → signal is saturated → no more proportional increase

might appear as indels (insertion/deletions)

Illumina – Sequencing by Synthesis (SBS):

Reversible termination (sanger has irreversible termination)

modified dNTP

dNTP – ter – fl

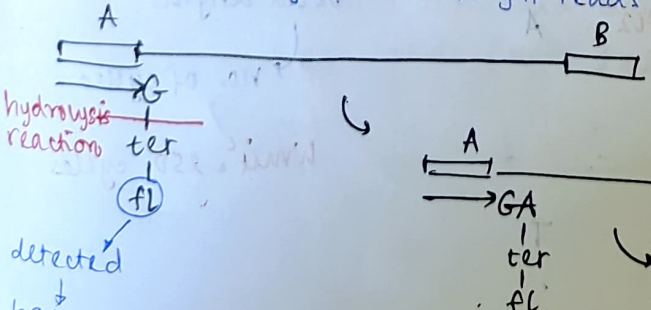
fl: fluorophore

ter: terminator

can be removed by a hydrolysis reaction

one base at a time

n cycles → n length reads

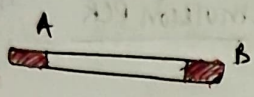


29/1/24

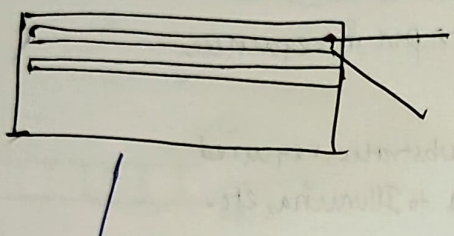
Illumina (SBS):

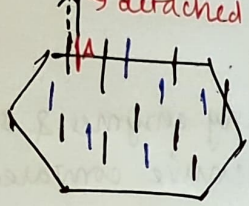
Reversible Termination

dNTP – ter-fl → diff. for the four bases: fl1, fl2, fl3, fl4

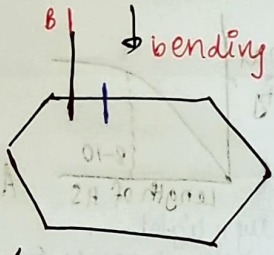


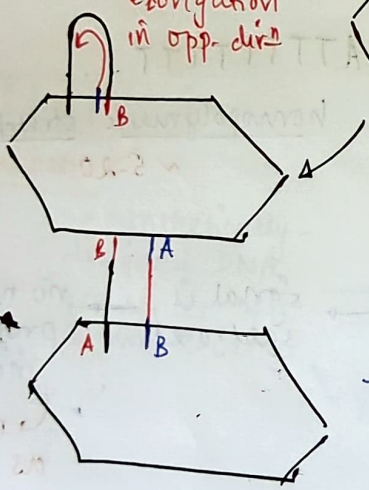
B → detached





Flow cell

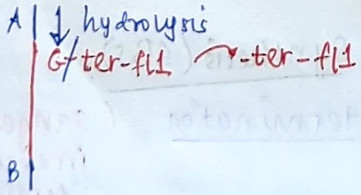




Bridge Amplification

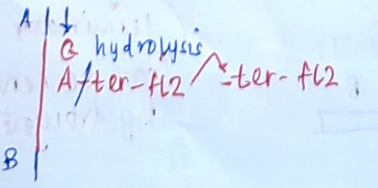
→ after many cycles, we get a cluster of one type of sample

camera



G

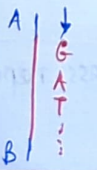
cycle 1



cycle 2

A

cluster has same bead length → no. of cycles → limit ~ 250 cycles



nth cycle

T

so on

in a cluster → ~100 molecules that are identical → detector reads average signal from a cluster

so after 250 ~~signals~~/bases, these avg. signals can be confusing & hard to decipher

in the adaptor seq:
- we add index sequences eg: ATTGCC
  - to identify sequences in a mixed sample

Multiplexing

Single-end (SE) and Paired-end (PE) sequencing:

↳ Illumina Data Processing.

↑ Throughput : high reads / lot of data

150 bp : SE seq.
2X150 bp : PE seq. (like in Illumina video).

Advantages:

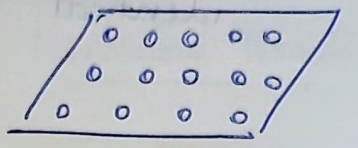
- cost-effective (not many enzymes are required).
- polymeric stretch → seq. one by one, not all at a time like Roche 454.
- highly accurate

Drawbacks:

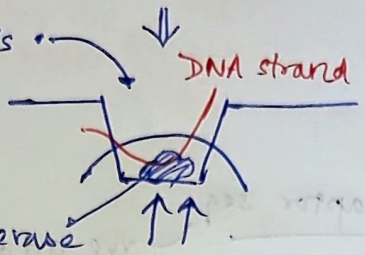
- expensive compared to the newer NGS ← even a μl costs in thousands.
- short reads (upto 2X250 bp)

Single Molecule Real Time Sequencing (SMRT): → [SBS]

↳ sequencing happens inside ZMWs.



SMRT-cell



everything else is similar to Illumina ←

immobilised at the bottom

light source → does not pass through, just illuminate the bottom

→ no amplification step is required.

→ no cluster formation

→ no averaging of signals → that's why 'single molecule'

↳ more sensitive detec[n]

as λ of light too large

signal detec[n] only at illumination

so incoming DNTPs do not produce signal

Advantages:

- long reads
- avg read length of 2–3 kb, some 10–20 kb
- single-molecule resolution

Drawbacks:

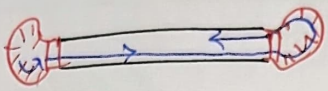
- high error rate (~14%)

Circular Consensus Sequence (CCS) reads → 99% accuracy

1. ATGCCTATT
2. ATCCCTATT
3. ATGCCAATT
4. ATGCATATT

take the most freq. at a place ↑

ATGCCTATT } Consensus sequence



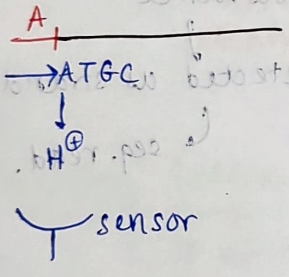
multiple repeats
↓
circular synthesis.

polymerase at the bottom → always illuminated by light → leads to adverse effect on fidelity
↳ high energy state

use of Illumina & SMRT together → Hybrid Sequencing
short reads but high accuracy
↳ large reads but low accuracy.

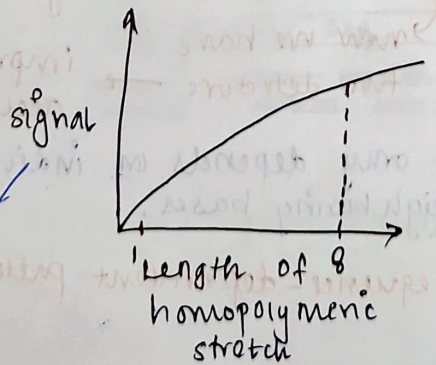
5|2|24

Ion Torrent Sequencing: → [SBS].



whenever there is a base addition, a hydrogen ion is released.

→ add one base at a time. → just like 454
→ add a dNTP → check change in $H^{\oplus}$ ion concentration.
→ same drawback as 454.



this drawback is related to the sensor

eg: A → T → G → C
↓
if an A is present in the template, a change in pH is observed at this stage.
↓
washed
↳ repeated.

Advantages:

- ✓ read length 400-700 bp
- ✓ accuracy ~ 99%
- ✓ no expensive optical detection unit required
- ✓ no modified dNTP required.

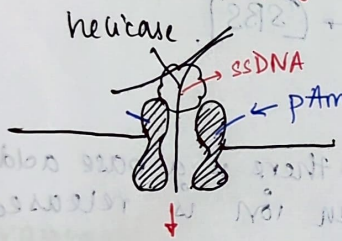
Drawbacks:

- ✓ lower throughput compared to Illumina

Oxford Nanopore Technology (ONT):

↳ Direct DNA Sequencing (not SBS).

→ no dNTPs
→ no polymerase
} → very economical.



✓ each base has different resistance → detected as strand passes → seq. read.

Advantages:

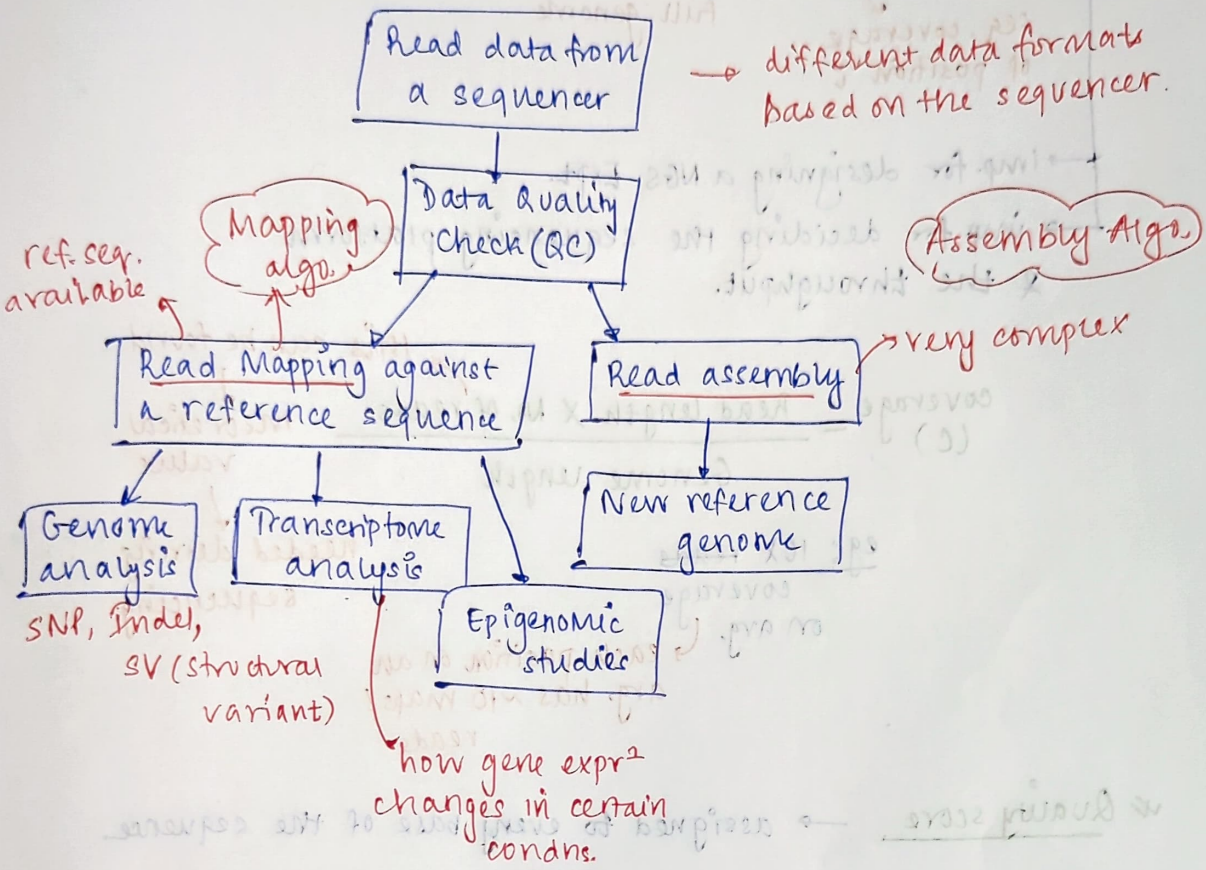
- ✓ long read seq. ~ 10-100 kb.
- ✓ real time observation

Drawbacks:

- ✓ ↑ error rate (~20%) → base detection from the current signal
  ↳ now we have two detectors → improved accuracy.
- ✓ current change not only depends on individual bases but also on the neighboring bases.
  ↳ sequence-dependent pattern

Data formats & Data quality:

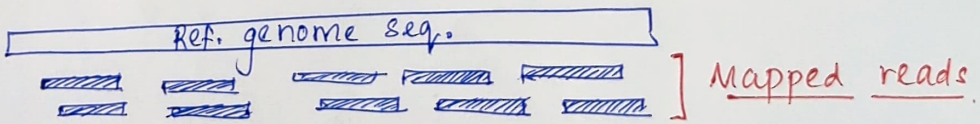
NGS Data analysis:



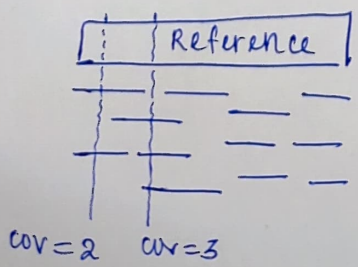
The Assembly Problem:

↳ if no ref. genome, how do we know where the genome starts from?

The Mapping Problem:



Sequencing Coverage / Sequencing Depth:



no. of reads mapped to that particular position

* **Mean coverage** → mean of coverage for all positions across the whole genome

$$= \sum_{i=1}^{L} cov_i / L$$

- cov_i → seq. coverage of position i
- L → length of the full genome

→ imp. for designing a NGS Expt.

→ imp for deciding the sequencing platform & the throughput.

$$\text{coverage } (C) = \frac{\text{Read length} \times \text{No. of reads}}{\text{Genome length}}$$

this can be found theoretical value needed during sequencing

eg: 10X coverage on avg. → each position on an arg. has ~10 maps/reads

* **Quality score** → assigned to every base of the sequence

12/2/24

FASTA
FASTQ → Illumina

SMRT
Ion Torrent
Nanopore

Read Mapping → the mapping problem is "string search"

Burrows-Wheeler Transform → Algorithm

PacBio SMRT seq. data format

HDF5 - Heirarchical Data Format version 5

BAM - a binary format → machine-readable → not human-readable
↳ can be converted to FASTQ
bam2fastx
↳ as most tools are based on this

Ion Torrent data format

raw data → DAT files - Windows format to store files
↳ pH change

seq. data → VCF -

~~HDF5~~

FAST5 data format

raw data
↓
procurement changes

analysis data
↓
sequence

Read Mapping:

Challenges:

- large genomes → human genome: 3×10^9 bp
- huge no. of short reads (millions or billions) — length of read $\sim 10^3$ bp
- time requirement → a long time for so many reads
- mutations & seq. error
- repetitive regions in the genome.
  ↳ a read can map to multiple regions if seq. is similar

Mapping Algorithms:

Data Structures:

1. Hash-table based mapping algorithms } very fast but take up a lot of memory. ↓ cannot be run on mere Desktop.
2. Suffix tree or suffix array based
3. Burrows-Wheeler transform based.
   - ref. genome only once → amt. of memory
   - very small

1 & 2 → stores the ref genome multiple times → ↑ memory

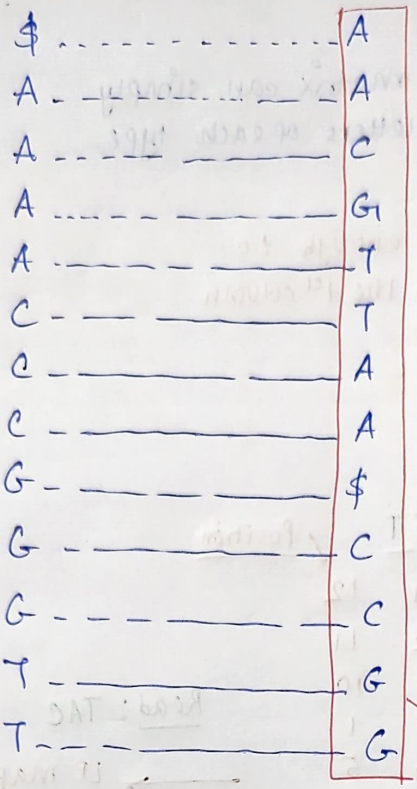
Burrows-Wheeler Transform (BWT):

Reference: G A C G T A C G T C A A $

STEP: 1

✓	$	G	A	C	G	T	A	C	G	T	C	A	A
✓	A	$	G	A	C	G	T	A	C	G	T	C	A
✓	A	A	$	G	A	C	G	T	A	C	G	T	C
✓	C	A	A	$	G	A	C	G	T	A	C	G	T
✓	T	C	A	A	$	G	A	C	G	T	A	C	G
✓	G	T	C	A	A	$	G	A	C	G	T	A	C
✓	C	G	T	C	A	A	$	G	A	C	G	T	A
✓	A	C	G	T	C	A	A	$	G	A	C	G	T
✓	T	A	C	G	T	C	A	A	$	G	A	C	G
✓	G	T	A	C	G	T	C	A	A	$	G	A	C
✓	C	G	T	A	C	G	T	C	A	A	$	G	A
✓	A	C	G	T	A	C	G	T	C	A	A	$	G

Alphabetical sort: ['$' comes before A]



STEP: 2

Burrows-Wheeler String

BWT(S) =
AACGTTAA$CCGG

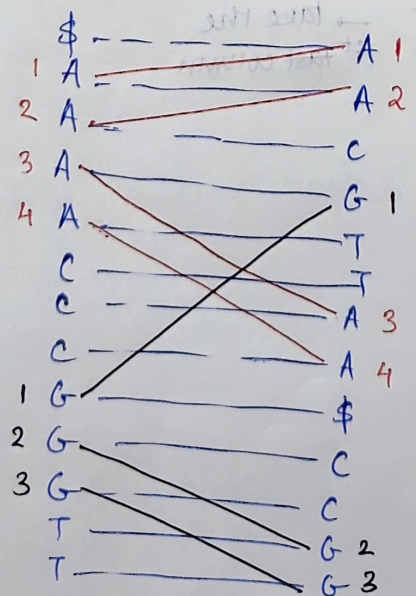
last column → sort alphabetically → first column

position info. also stored.

(say) search read ACGT → how do we take into account mismatches? AGGT is closest to ACGT but with one mismatch

Last-first (LF) mapping:

the i^{th} occurence of character c in the last column & the i^{th} occurence of character c in the first column correspond to the same character in the ref. seq.



if we no. the As in first column & no. the As in last column, the 1-1, 2-2, 3-3, ..., n-n correspon & occupy the same positio in the reference genom

FM Index:

Full-Text Minute-space (Ferragina, Manzini)

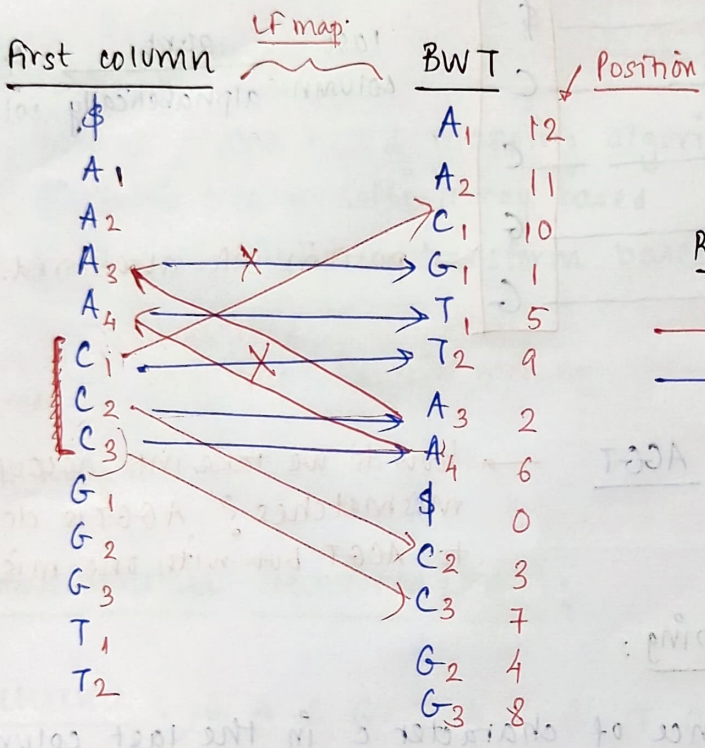
↓

first column of the BWT matrix can simply be stored as numbers of letters of each type.

A	C	G	T
4	3	3	2

this is enough to generate the 1st column

Searching a read position:



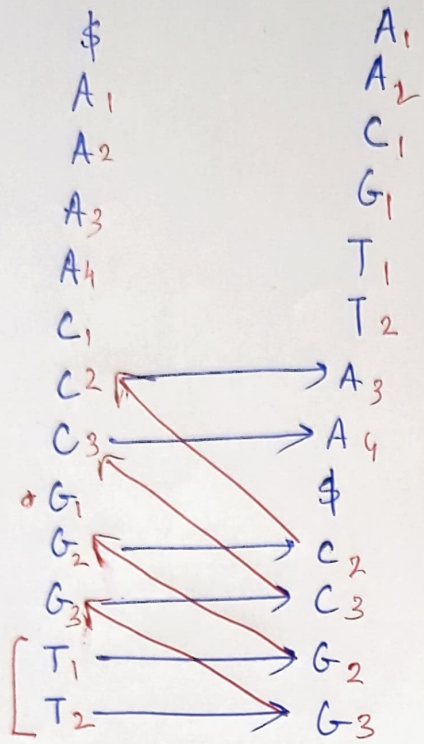
Read: TAC

→ LF mapping strings that come before C in seq.

Read: T A C

we have an A coming

start from the last → take the 1st ~~last~~ column



ACG[T]

both ways/tracks give us ACGT

repeat sequences