BMEG 3105 Fall 2025

Data analytics for personalized genomics and precision medicine

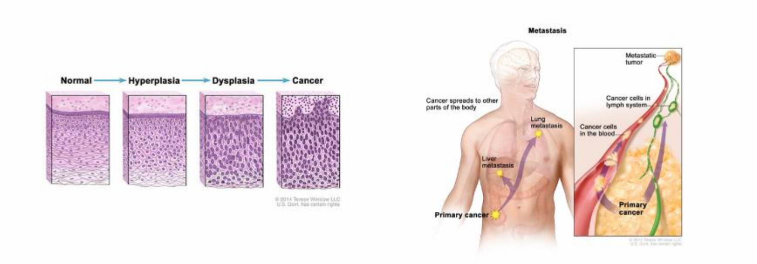
Lecturer: Yu LI (李煜) from CSE

LECTURE 15: Cancer genomics overview &genomics analysis (31/10/2025)

Scriber: Rana Sabri (1155228843)

What is Cancer?

Cancer is a disease in which some of the body’s cells grow uncontrollably and spread to other parts of the body.

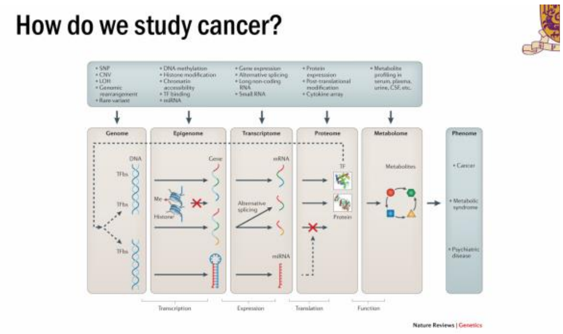


Why do we want to study Cancer?

Cancer, after cardiovascular diseases is the second disease that takes the greatest number of lives. Half of us is likely to be diagnosed with cancer in our lifetimes, so it’s crucial that we continue to study and learn about this complex illness so that we can stop it from cutting lives short all over the world. (Cancer Research UK, 2023)

How do we study cancer?

Cancer is usually believed to be a genomic disease so we will use genomics/multi-omics methods to study it, by looking at Genome/Epigenome/Transcriptome/Proteome/Metabolome.



How do we study cancer?

QUESTION; Which would not be affected by epigenomic modification?

1. DNA sequence
2. Gene expression (epigenetic changes can turn genes on and off)
3. The transcription process (these changes can make DNA more or less like accessible to transcriptome.

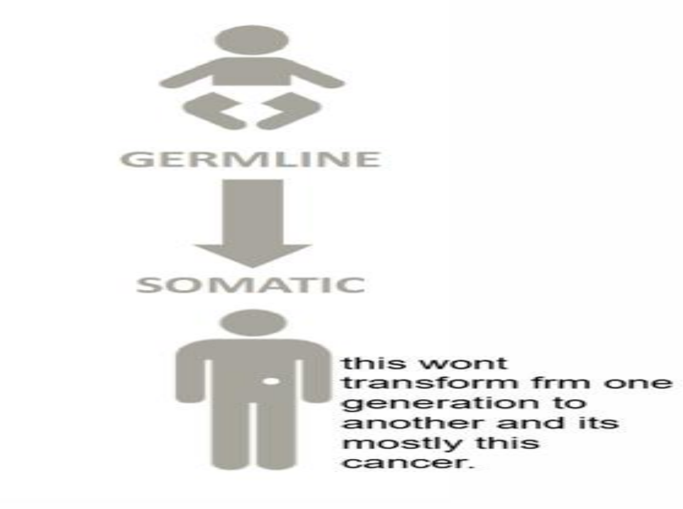
4. Disease and phenotype (epigenetic changes lead to diseases like cancer)

Variants

Why do we care about variants? ➔ Almost 99.5% of our genome is similar to other people, that is why we need to look at the variants in order to determine other diseases that separate us from the remaining population.

Different types of genomic variants

1) short variation (<50bp) only a change for a single nucleotide
2) CNV =the entire gene could have multiple copies
3) SV= part of entire chromosome different
4) PathSeq= it is a non-human sequence for example when virus gene could be incorporated in the human dna sequence



this wont transform frm one generation to another and its mostly this cancer.

How to discover the genetic variants?

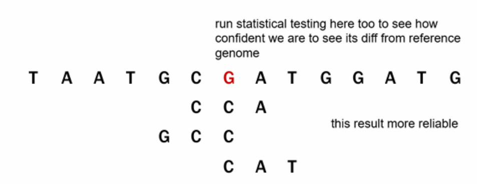
a. Library preparation b. Sequencing

Sequence mapping recap

a. Slide each read along the genome, calculate the difference

However we should take multiple reads to have more confidence and accuracy

For example like this;



run statistical testing here too to see how confident we are to see its diff from reference genome

```
T A A T G C G A T G G A T G
          C C A
        G C C
            C A T
```

this result more reliable

Variants VS errors

Sometimes we may think errors could be actual variations, therefore we need to be able to differentiate between them.

Errors can creep in on various levels:

➢PCR artifacts (amplification of errors)

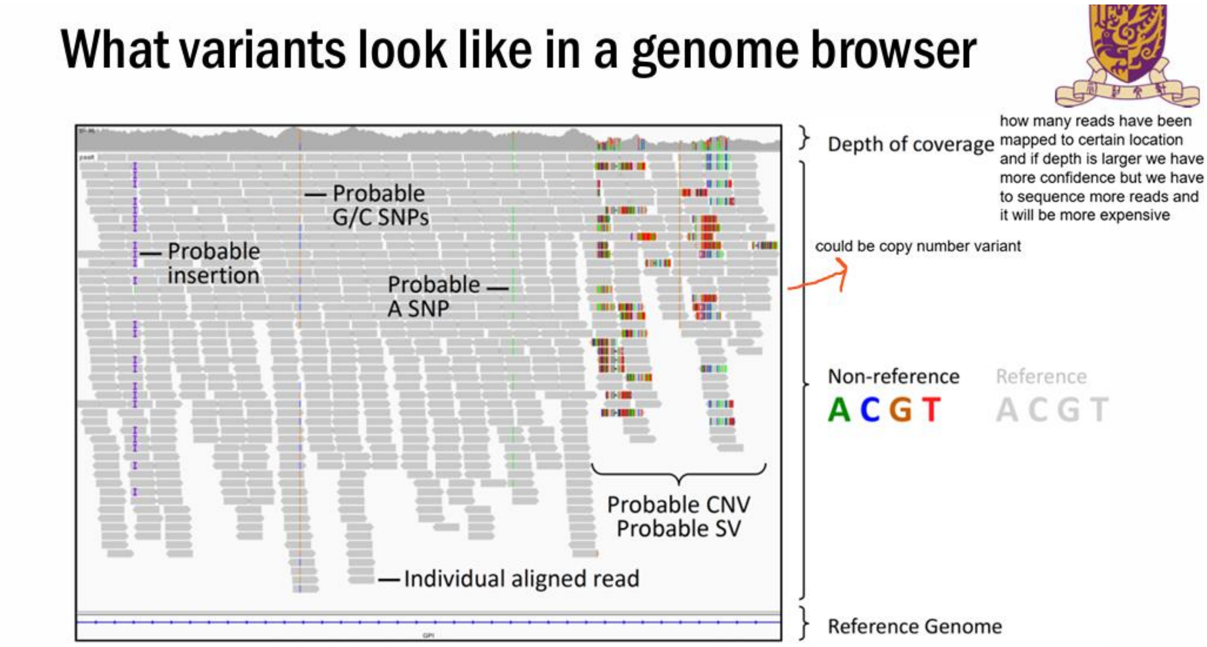
➢Sequencing (errors in base calling)

➢Alignment (misalignment, mis-gapped alignments) ➢Variant calling (low depth of coverage, few samples)

➢Genotyping (poor annotation)

Also for sequencing, Illumina sequencing have a higher accuracy rate (99.999%) compared to nanopore sequencing which is around 99.5%, both seem high however we have millions if not billions of data in front of us so that could lead to thousands of important errors which we don't want, especially in cancer screening.

THIS IS WHAT WE ARE LOOKING AT



What variants look like in a genome browser

how many reads have been mapped to certain location and if depth is larger we have more confidence but we have to sequence more reads and it will be more expensive

could be copy number variant

Data pre-processing step

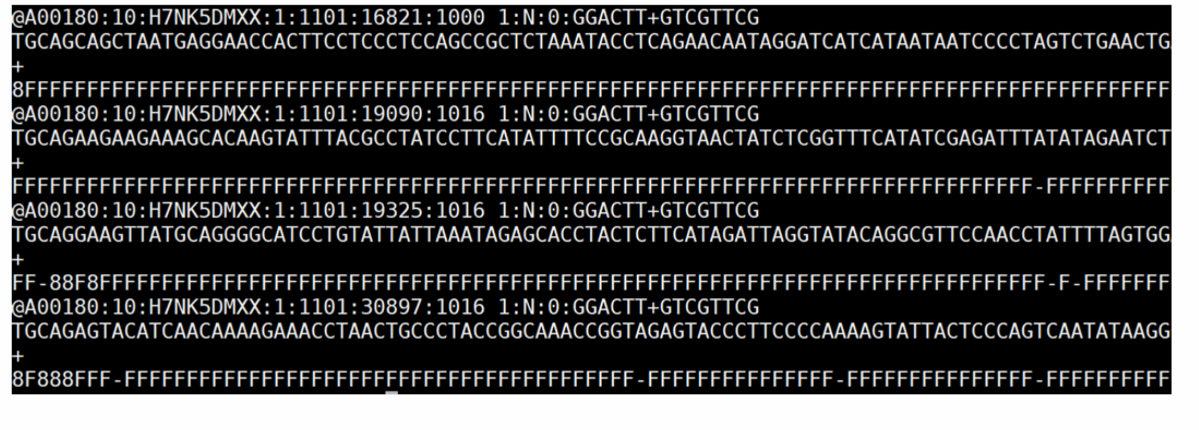
a. Mapping
b. i. Map the reads produced by the sequence to the reference
c. ii. Mapping and alignment algorithms: BWA for DNA, STAR for RNAseq
   iii. Input format: FASTQ
d. iv. Output format: Sequence/ Binary Alignment Map (SAM/ BAM)

e. v. CIGAR (Concise Idiosyncratic Gapped Alignment Report)

b. Marking duplicates

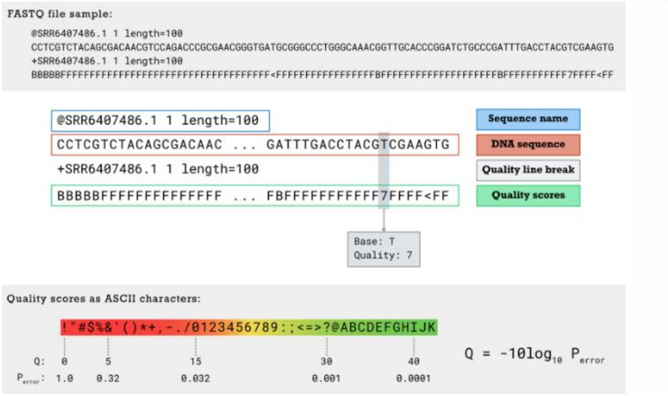
i. Duplicates: non-independent measurements of a sequence fragment= We might think duplicated could mean cancer mutations however sometimes it means its just artefacts, it could usually come from PCR or optical duplicates from screening.

ii. Must be removed to assess support for alleles correctly



```
@A00180:10:H7NK5DMXX:1:1101:16821:1000 1:N:0:GGACTT+GTCGTTCG
TGCAGCAGCTAATGAGGAACCACTTCCTCCCTCCAGCCGCTCTAAATACCTCAGAACAATAGGATCATCATAATAATCCCCTAGTCTGAACTG
+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG
+
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFF
```

THIS IS FASTQ, HERE 4 LINES CORRESPONDS TO ONE READING



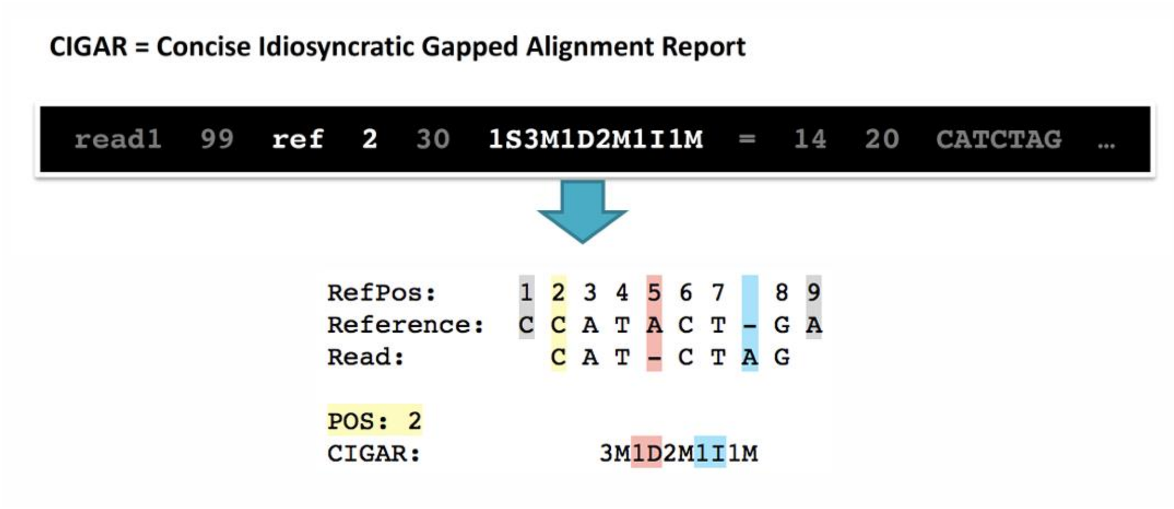
FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTGCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFF7FFFF<FF
```

@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAAC ... GATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFF ... FBFFFFFFFFFFF7FFFF<FF
Sequence name
DNA sequence
Quality line break
Quality scores
Base: T
Quality: 7
Quality scores as ASCII characters:
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK
Q: 0 5 15 30 40
P_{error}: 1.0 0.32 0.032 0.001 0.0001
$Q = -10\log_{10} P_{error}$

HERE CORRESPONDS TO LOW ACCURACY HERE CORRESPONDS TO HIGH ACCURACY

SAM and BAM have same content only difference is that humans can read text stored in SAM and computers can read the binary code stored in BAM.



CIGAR = Concise Idiosyncratic Gapped Alignment Report

```
read1   99   ref   2   30   1S3M1D2M1I1M   =   14   20   CATCTAG   …
```

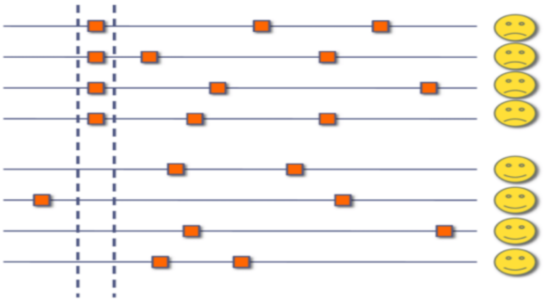
```
RefPos:      1 2 3 4 5 6 7   8 9
Reference:   C C A T A C T - G A
Read:          C A T - C T A G

POS: 2
CIGAR:           3M1D2M1I1M
```

This is CIGAR, where there is either a (mis)match, insertion, deletion.

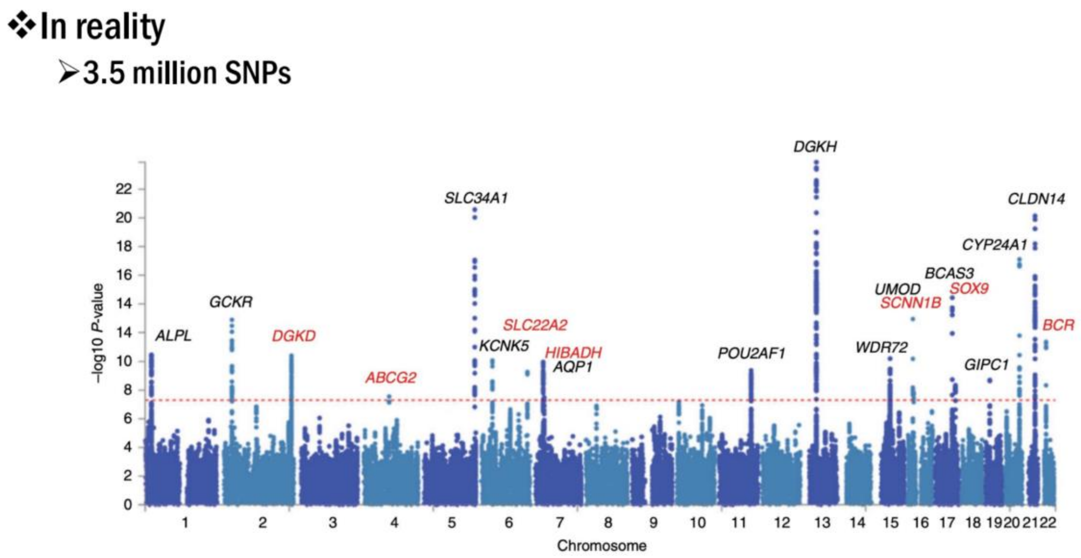
GWAS

Try to determine whether specific variant(s) in many individuals can be associated with a trait (disease).



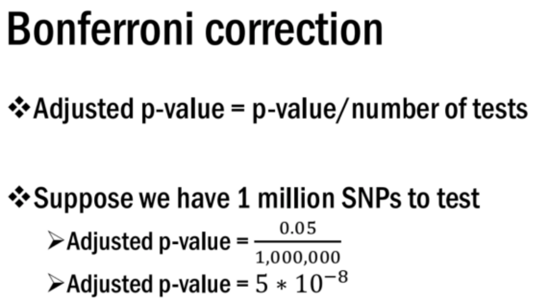
These could be the variants associated with the disease.

This method is ideal for some rare Mendelian diseases as they are not complex diseases and only one or two variants cause the diseases.



- ❖In reality
  - ➢3.5 million SNPs

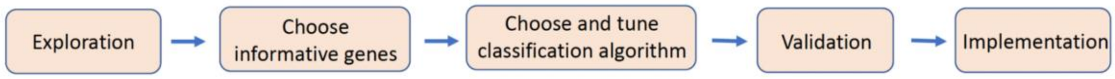
This is usually the cases that we deal, if we keep in mind that there is a p=0.05% of each snippet out of 3.5 million snippets (which are independent), then 1-(0.95)^3.5million, which is around 1, shows that there will unfortunately be an error throughout these snippets, so to reduce the errors we should reduce the error rate of each snippet making the confidence rate higher.



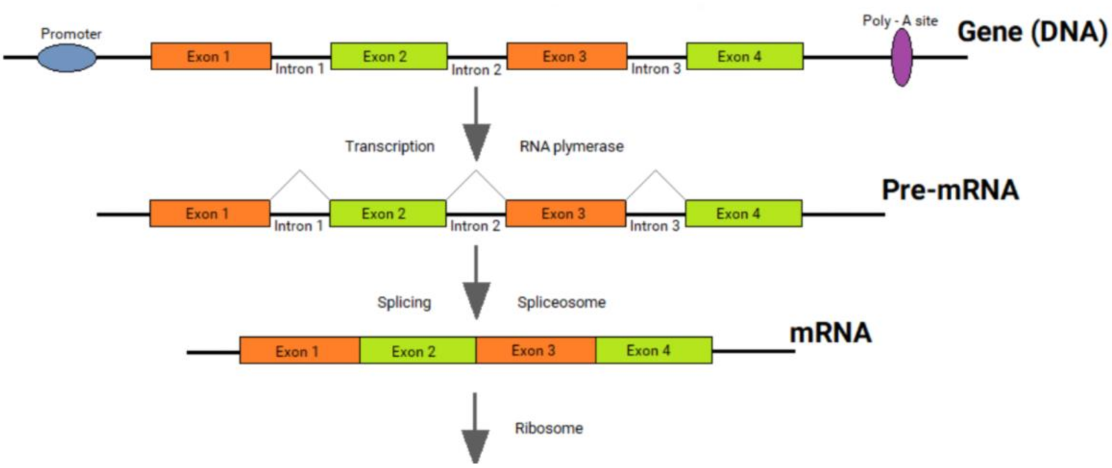
Bonferroni correction

- ❖Adjusted p-value = p-value/number of tests
- ❖Suppose we have 1 million SNPs to test
  - ➢Adjusted p-value = $\frac{0.05}{1,000,000}$
  - ➢Adjusted p-value = $5 * 10^{-8}$

RNA-seq data analysis

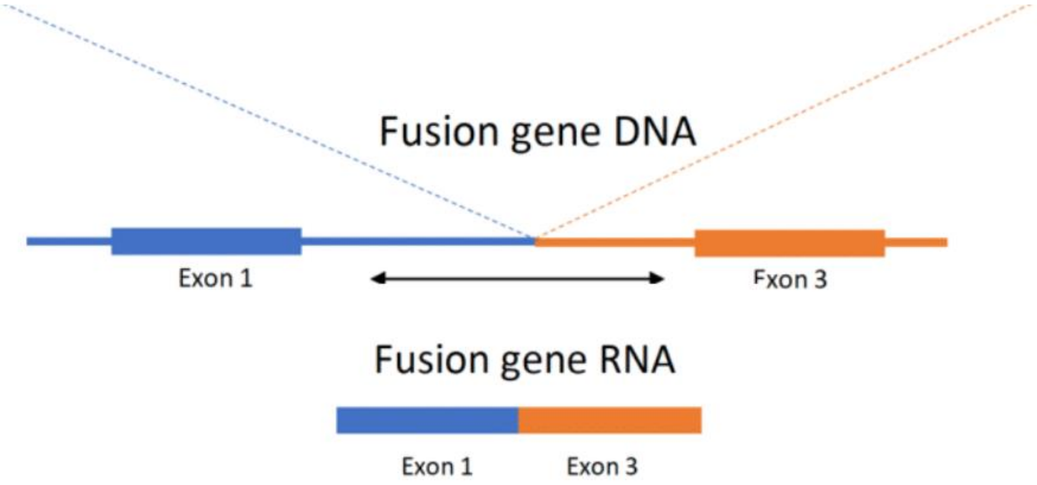


Exploration → Choose informative genes → Choose and tune classification algorithm → Validation → Implementation

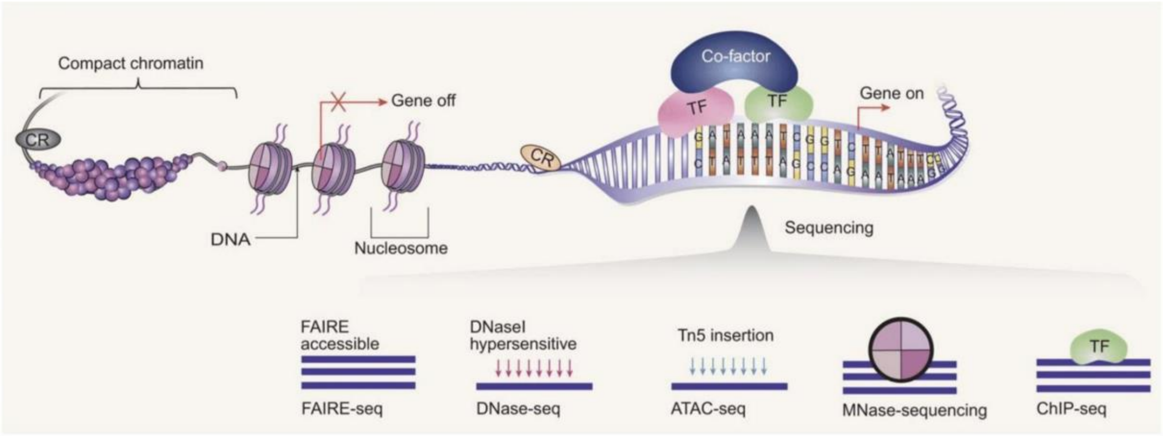


As shown above we can see the splicing of the eukaryotic gene where introns are being cut and the remaining exons are joined, however Break-points are in introns and we need whole genome sequencing as whole exome sequencing is not enough.

This process is also an expensive process, therefore there is another process called GENE FUSION; which is more convenient.



Epigenomics and Sequencing protocols



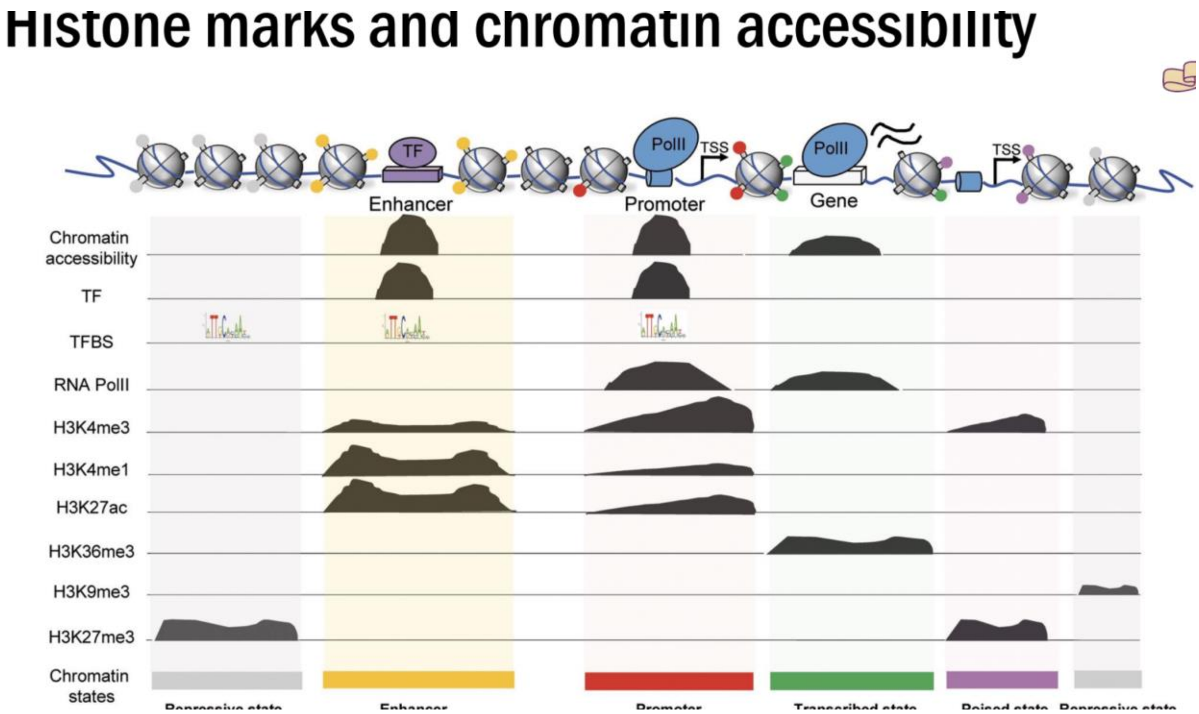
Studying how DNA is wrapped around nucleosomes

Identifying where specific proteins (like transcription factors) bind to DNA.

The overall data analytics pipeline for epigenomics

1) sample preparation and sequencing
2) computational analysis

Browser Extensible Data (BED) ➢Chromosome➢Start➢End➢Label➢...



Histone marks and chromatin accessibility

Here we can see that the gene is tightly held which makes it not accessible.

Enhancers show peaks for H3K4me1 and H3K27ac, meaning they boost gene activity from a distance. Promoters have peaks for H3K4me3, PolII, and TFs, showing they start transcription. Gene regions show PolII peaks, meaning transcription is happening.

REFERENCE

[1] Li, Yu (2024). BMEG3105: Data Analytics for Personalized Genomics and Precision Medicine - Cancer Genomics Overview & Genomics Analysis.

[2] **Cancer Research UK. 2023.** “About Our Information: Sources and References.” *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancer/about-our-information/writers-guidelines/sources-references.