BMEG 3105 Fall 2023

Data analytics for personalized genomics and precision medicine

Topic: Genomics data analysis

Lecture: Lecturer: Yu LI (李煜) from CSE

Liyu95.com, liyu@cse.cuhk.edu.hk

Student: CHANG Hing Lam SID: 1155143887

4th November, 2023

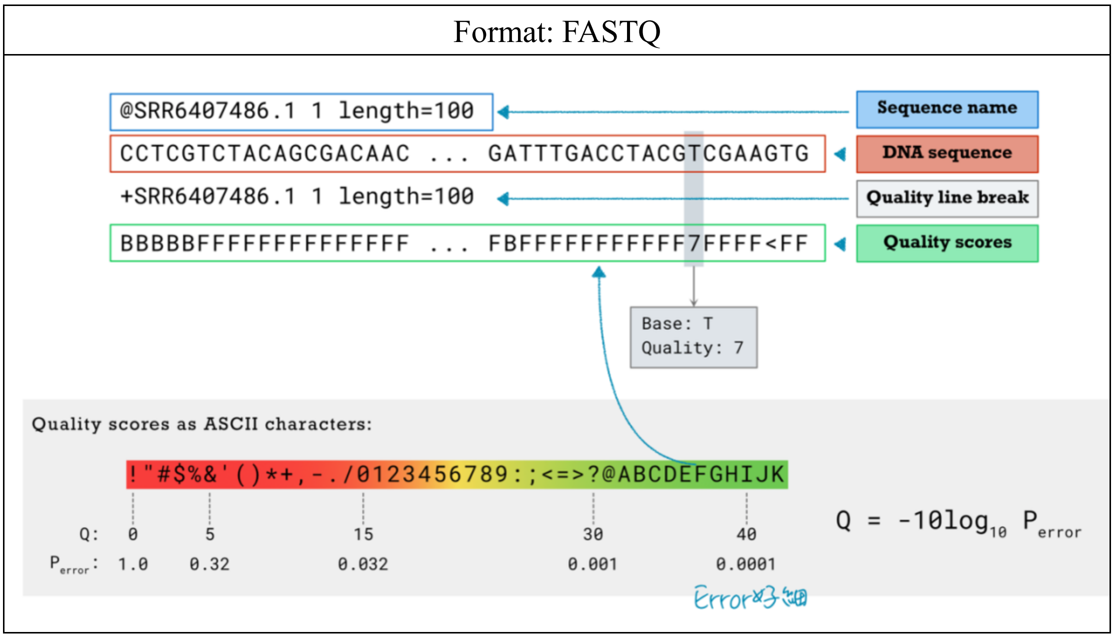
Expected Outcome

1. Variant calling pipeline
   - Understanding reasons for the steps, file interpretation and factors affect variant calling.
2. Gene fusion
   - Understanding the definition and RNA-seq.
3. GWAS
   - P-value correction
4. Epigenetics
   - Understanding gene expression regulation: structure and environment, and Data analytics pipeline

Data Preprocess Step

Raw Unmapped Reads

Format: FASTQ



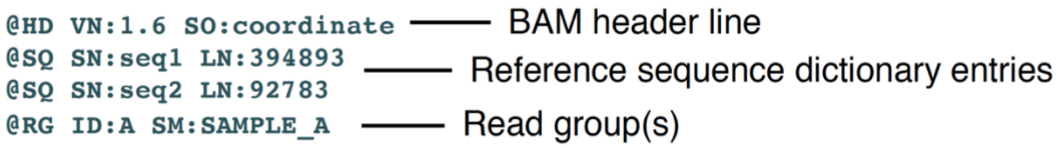
```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAAC ... GATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFF ... FBFFFFFFFFFFF7FFFF<FF
```

1. Mapping
    - BWA for DNA
    - STAR for RNAseq

Raw Mapped Reads

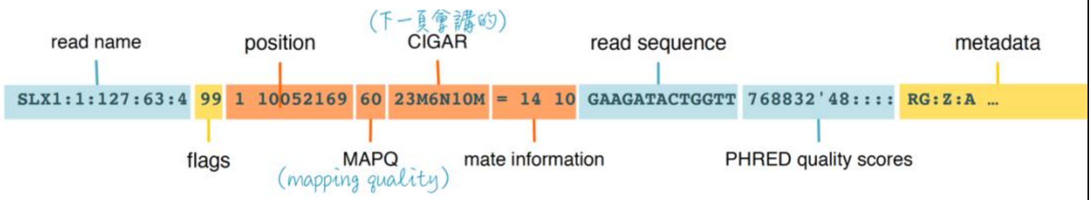
Format: SAM / BAM

Header Line:



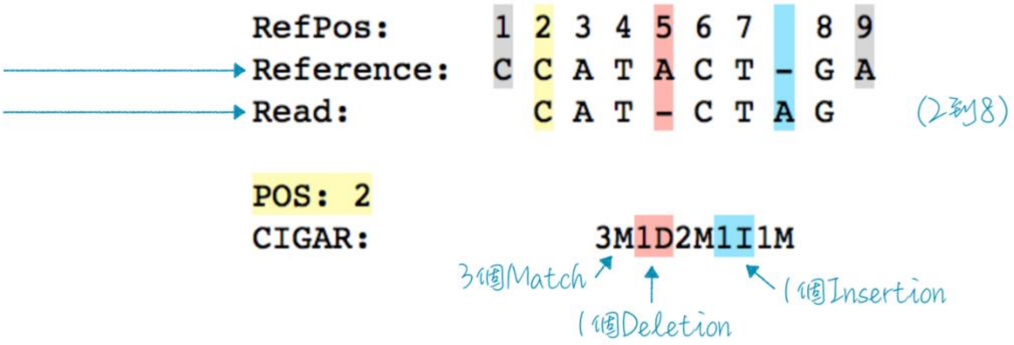
```
@HD VN:1.6 SO:coordinate
@SQ SN:seq1 LN:394893
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A
```

Records:



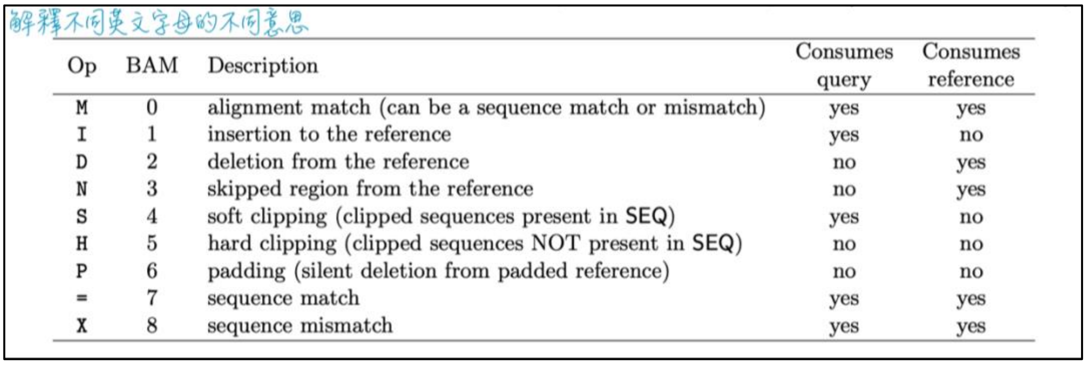
```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: RG:Z:A …
```

CIGAR



```
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
```

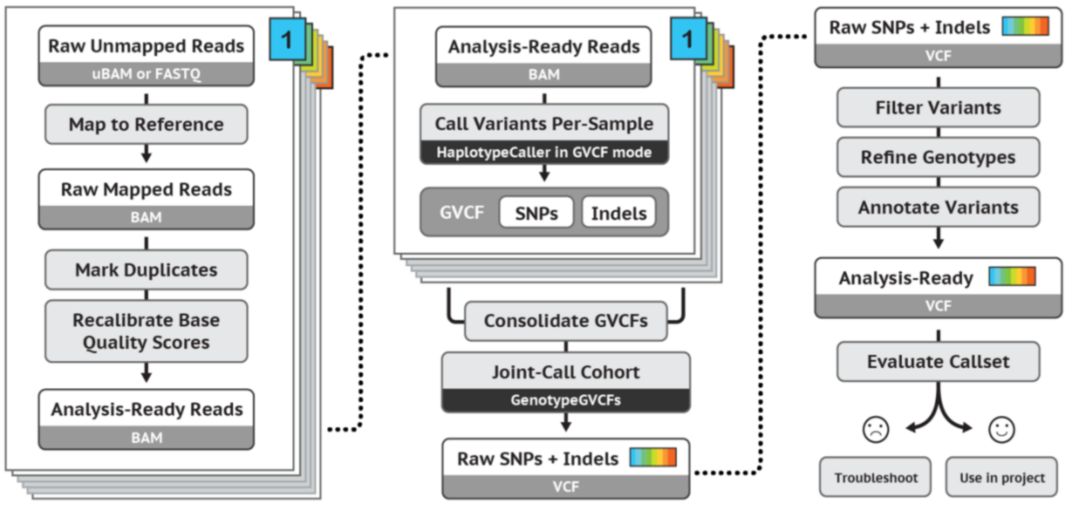


解釋不同英文字母的不同意思

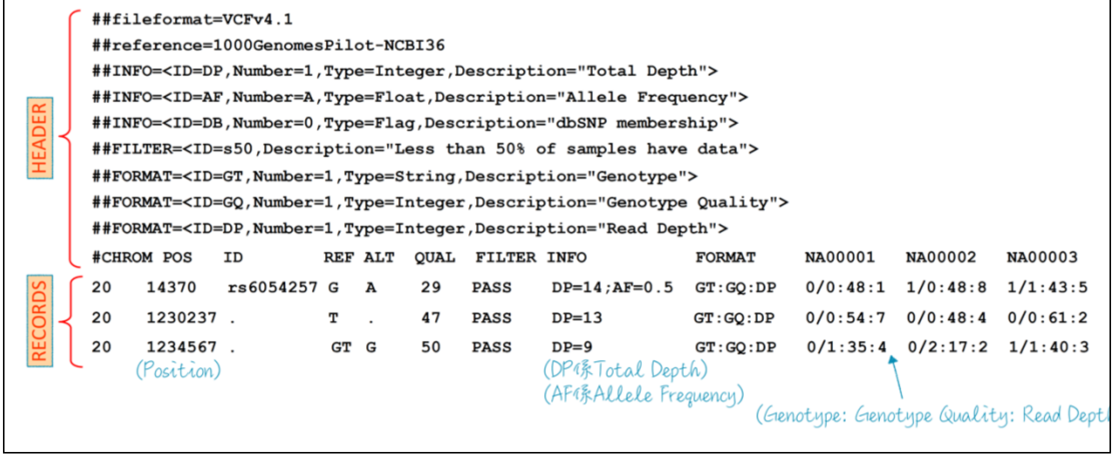
Op	BAM	Description	Consumes query	Consumes reference
M	0	alignment match (can be a sequence match or mismatch)	yes	yes
I	1	insertion to the reference	yes	no
D	2	deletion from the reference	no	yes
N	3	skipped region from the reference	no	yes
S	4	soft clipping (clipped sequences present in SEQ)	yes	no
H	5	hard clipping (clipped sequences NOT present in SEQ)	no	no
P	6	padding (silent deletion from padded reference)	no	no
=	7	sequence match	yes	yes
X	8	sequence mismatch	yes	yes

2. Marking Duplicates
    - Library Duplicates
    - Optical Duplicates

Variant calling



Variant Call Format (VCF)



```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM POS     ID        REF ALT  QUAL  FILTER INFO         FORMAT    NA00001   NA00002   NA00003
20     14370   rs6054257 G   A     29   PASS   DP=14;AF=0.5 GT:GQ:DP  0/0:48:1  1/0:48:8  1/1:43:5
20     1230237 .         T   .     47   PASS   DP=13        GT:GQ:DP  0/0:54:7  0/0:48:4  0/0:61:2
20     1234567 .         GT  G     50   PASS   DP=9         GT:GQ:DP  0/1:35:4  0/2:17:2  1/1:40:3
```

(Position)

(DP係Total Depth)
(AF係Allele Frequency)

(Genotype: Genotype Quality: Read Dept

Joint analysis

per-sample GVCFs → Final multi-sample VCF

[The value of QULAL increase]

3. Base Recalibration

What Final is going to Test:

1. Reasons that we need to do the steps.
2. Ability to read the records in those files.
3. How different factors affect the quality of the mapping and the variant calling.

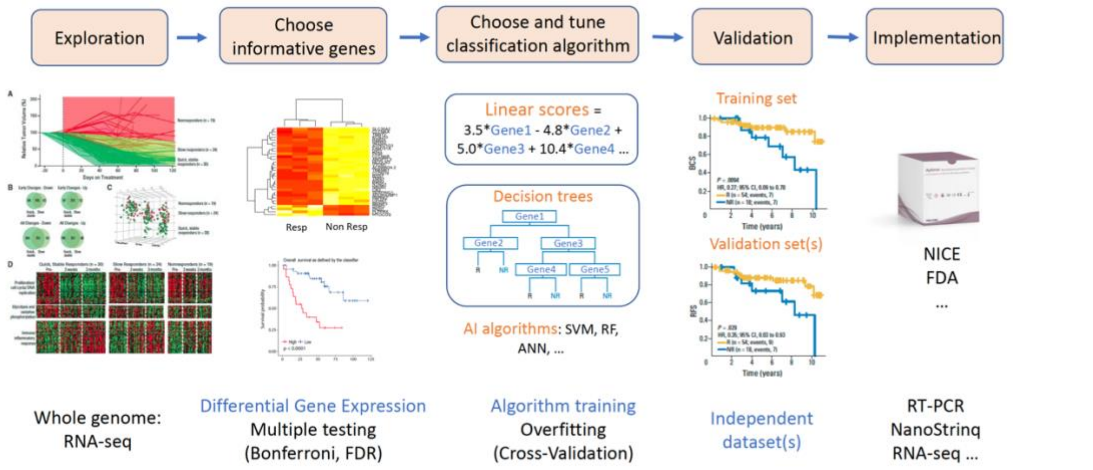
Genome-Wide Association Studies (GWAS)

- Spot the variant that is common amongst all affected

Bonferroni correction

- Adjusted P-Value = P-Value / Number of Tests

RNA-seq data analysis



Gene-Fusion

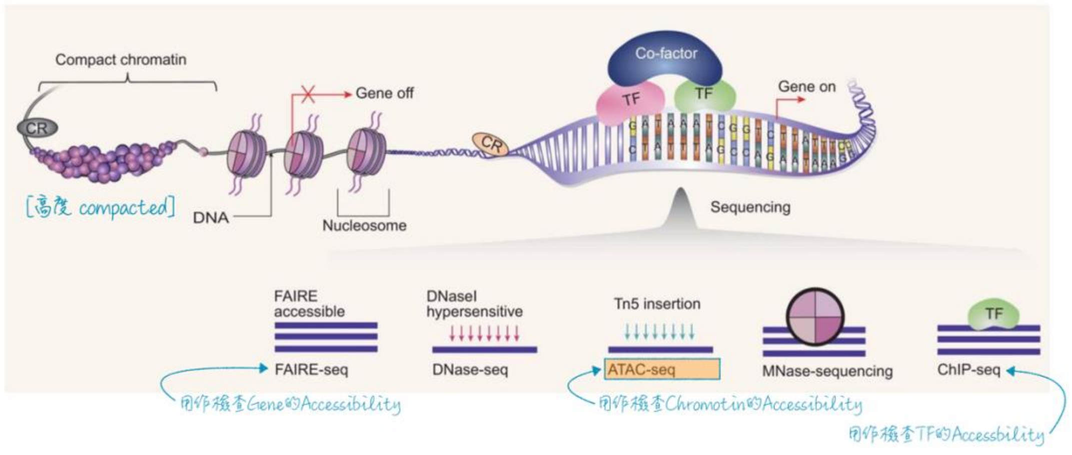
[Chromosomal Translocation, Interstitial Deletion, Chromosomal Inversion]

- Discovered in cancer cell in 1980s.
- Formed by fusion of two distinct wild type genes.
- Produced by somatic genome rearrangements in cancer.
- Required whole genome sequencing.

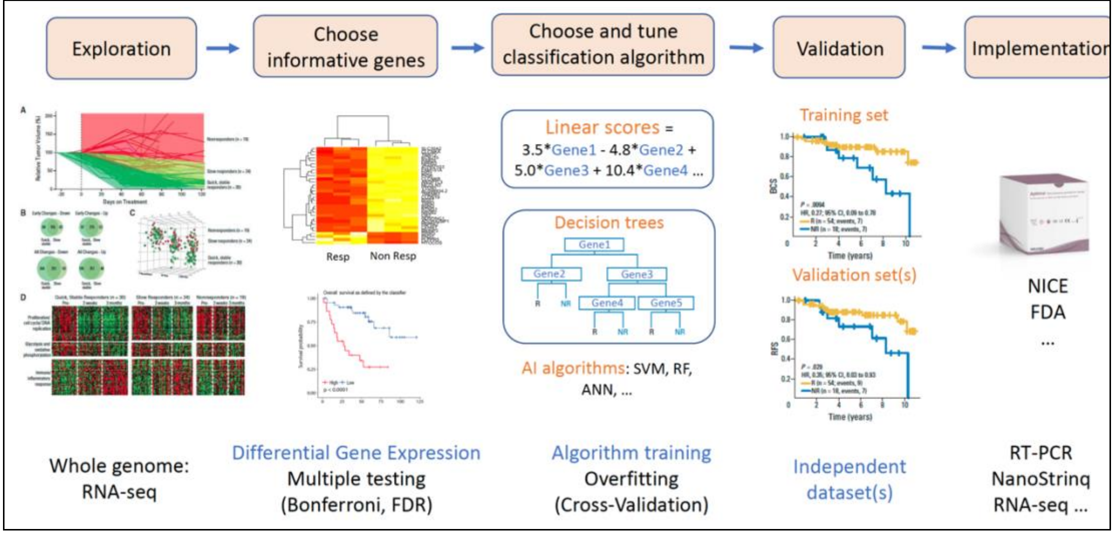
Abnormal gene expression

- Epigenetics

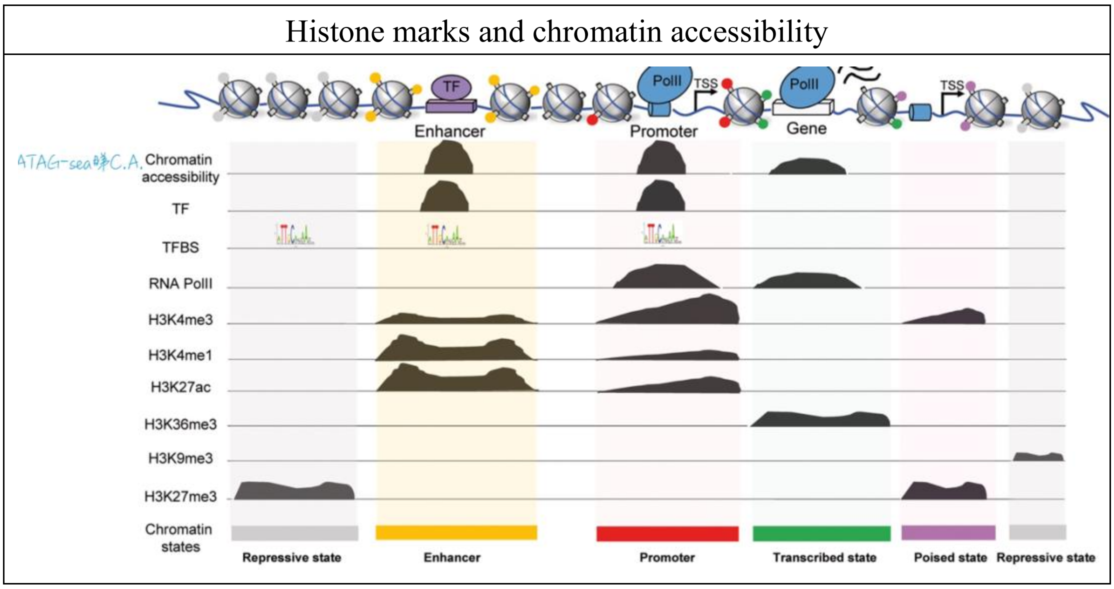
Sequencing protocols



Peak Calling [Statistical testing: Peak shape VS random background]



Histone marks and chromatin accessibility



Final Take Home Message:

No need to understand the "entire detailed pipeline", focus on the understand of Epigenetics, Sequencing Process, and Peak Calling Process.

Resources

https://www.ebi.ac.uk/training/materials/cancer-genomics-materials/

GATK workshop slides:

https://drive.google.com/drive/folders/1y7q0gJ-ohNDhKG85UTRTwW1Jkq4HJ5M3

GATK workshop video: https://www.youtube.com/watch?v=sM9cQPWwvn4

GWAS workshop:

https://www.youtube.com/watch?v=xw419NKqMqw

Epigenetics: https://www.youtube.com/watch?v=IAu44BkOaSs

https://www.encodeproject.org/atac-seq/