BMEG3105 Lecture 17

WU Sio Fong 1155173201

1.Cancer

- ❖ Definition of cancer:
  Disease in which some of the body's cells grow uncontrollably and spread to other parts of the body
- ❖ How to study cancer?
  - Genetic variants
    - Genome
    - Gene fusion (RNA-seq)
  - Abnormal gene expression
    - Genome (genetic information)
    - Epigenome (environment)
    - Transcriptome (direct measurement)

2.Overview of today lecture

- ❖ Genome
  - Variant calling
  - GWAS
- ❖ RNA-seq
  - Gene fusion---structural variant
- ❖ Epigenome
  - Peak calling

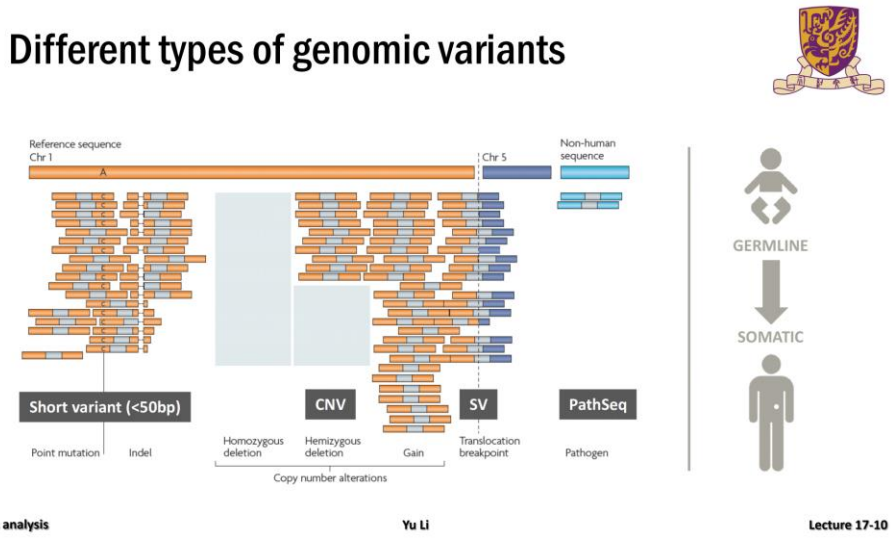
3.Genome

- ❖ **Variant calling**

  Reason:
  - 3.2 billion sites in the human genome
    • Any 2 humans share 99.5% DNA
    • Can efficiently describe a genome with relation to a reference
  - Genetic differences can lead to differences in disease risk and response to treatment
  - Genetic variation can used to find genes and variants that contribute to disease
  - Cancer: genetic variants at multiple levels

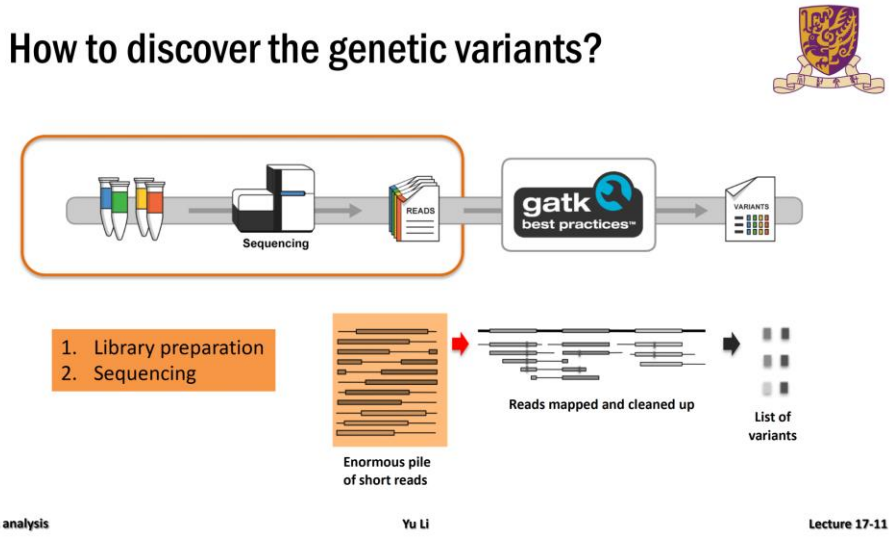
  Types of variant calling:
  - Short variant: point mutation, indel(<50bp)
  - CNV: homozygous deletion, hemizygous deletion, gain

- SV: translocation breakpoint (gene shift from other location)
- PathSeq: pathogen (non-human)



Ways to discover genetic variants:

Library preparation and sequencing



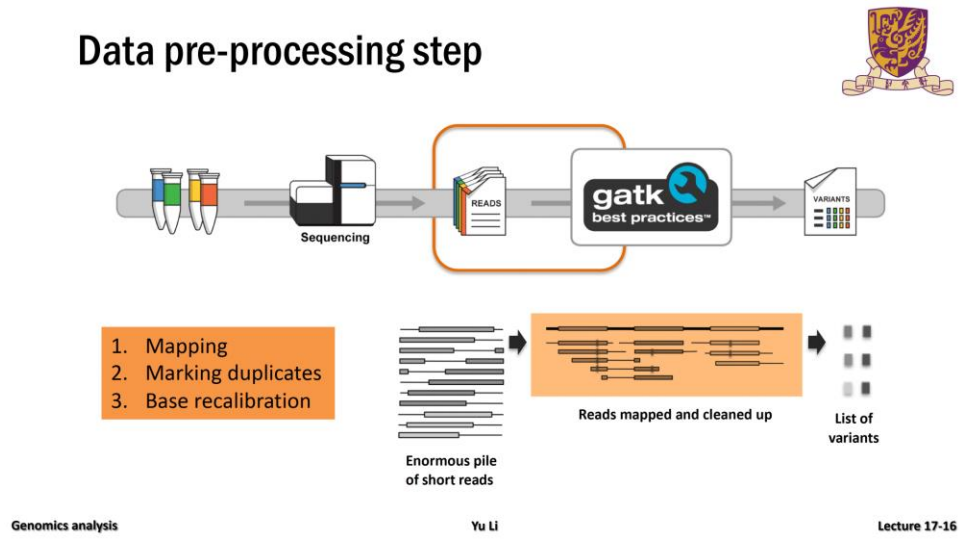
Variants VS errors

actual variation (real change)

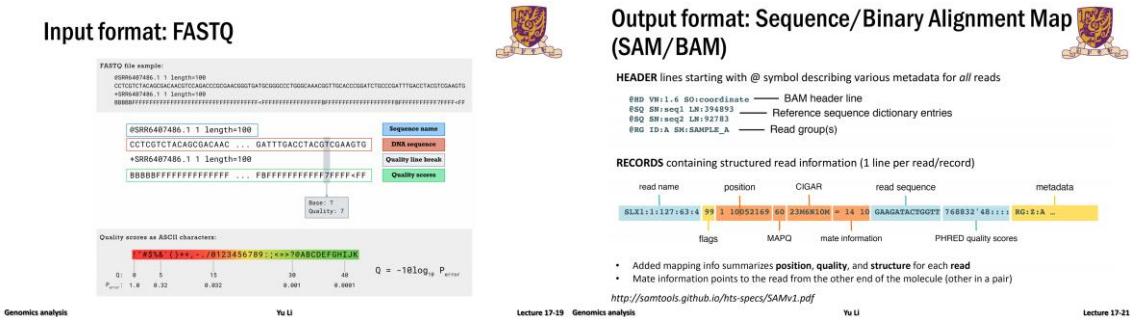
errors (artifacts) -- Errors creep in on various levels

- 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)

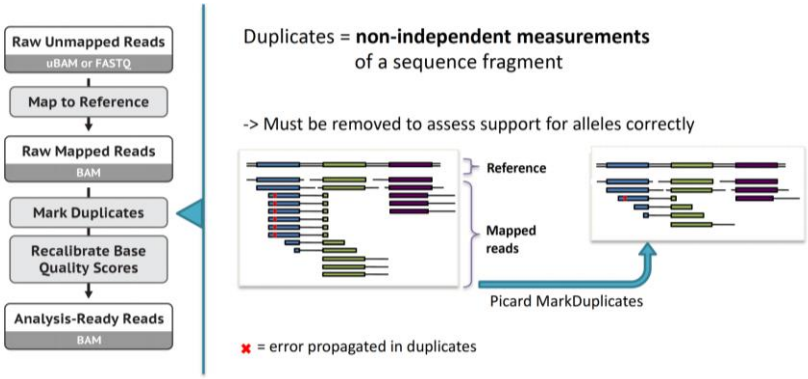
Procedure for data pre-processing:



Step1: map the reads produced by the sequence to the reference
Input: FASTQ, a text-based format for storing both a biological sequence and its corresponding quality scores.
Output: SAM/BAM.
SAM: Sequence Alignment Map (store in text)
BAM: Binary Alignment Map (story in binary only i.e., 1& 0)
The following pictures are FASTQ and BAM format respectively.



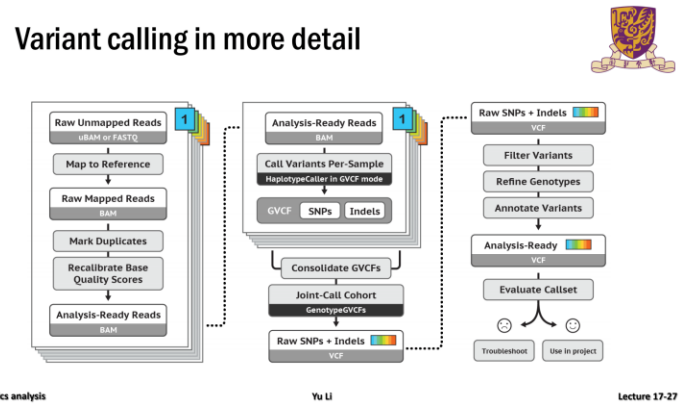
Step2: mark duplication to reduce duplications which come from some experimental manipulations.



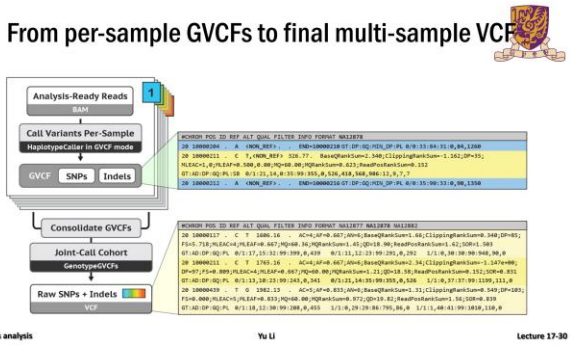
Error may be due to noise/ error/ duplicate (we must do this to avoid we mistake error as real variance)

Variant calling:

After analysis of data from the above operations, we want to find variants from reads, so we do:



VCF is some information in below form:

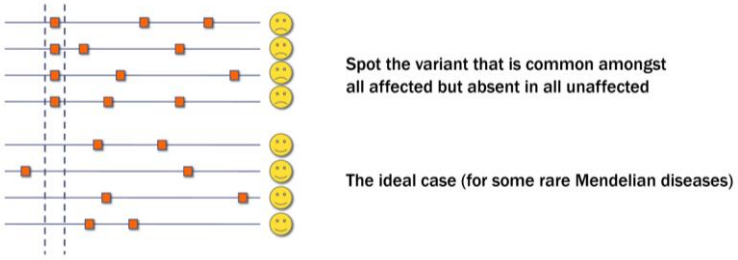


However, after analysis of genomes, we should **joint some data together** to conclude a result because a single genome data is always unpowered, a joint call set could provide more valuable information.

4.Further downstream analysis for cancer diseases:

GWAS:

Trying to determine whether specific gene variant is related to a disease

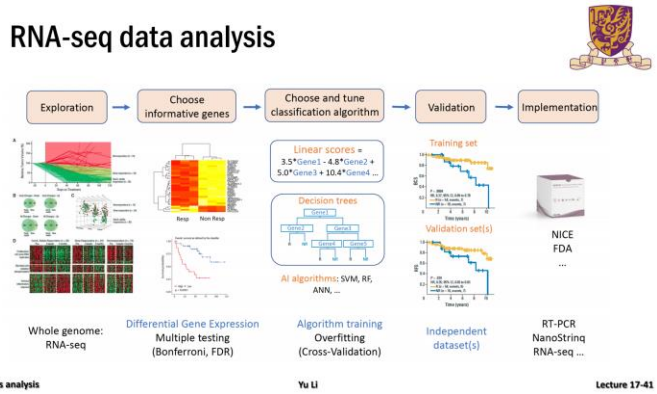


.

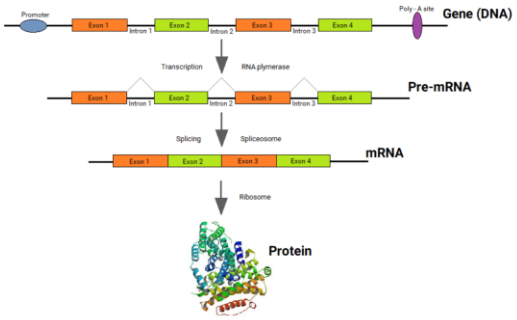
To find the correlation between SNPs to diseases:
Bonferroni correction --adjust P-value
Adjusted p-value= p-value/ number of tests

RNA-seq data analysis (genetic variant level to study cancer)

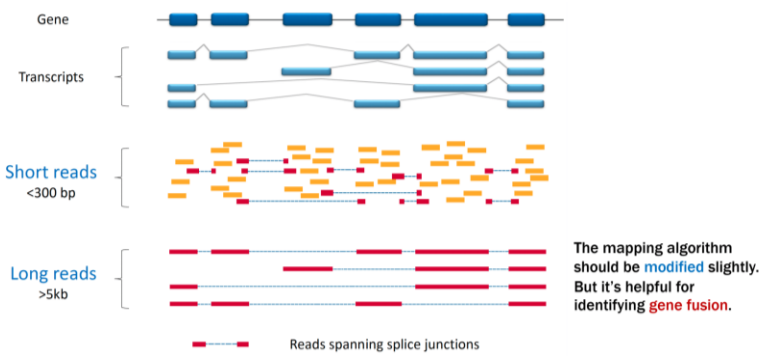
A basic procedure of RNA-seq data analysis



Transcription, splicing and translation of a eukaryotic gene



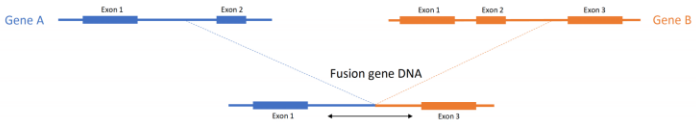
Mapping spanning splice junctions



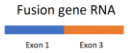
Gene fusion:

- Novel gene formed by fusion of two distinct wild type genes
- Is a specific kind of structural variant related to cancer
- In cancer: produced by somatic genome rearrangements

RNA-seq for gene fusion detection



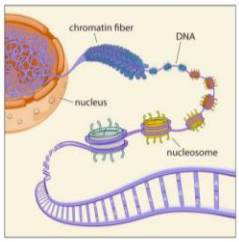
Break-points are in introns
We need whole genome sequencing
Whole exome sequencing is not enough

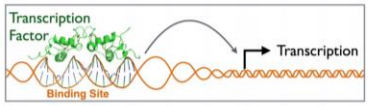


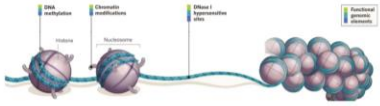
Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads

5.Epigenome

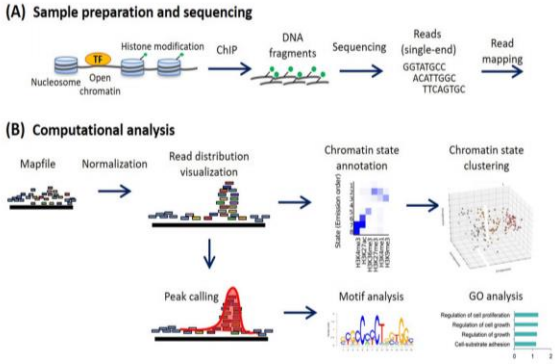
Structure of epigenome



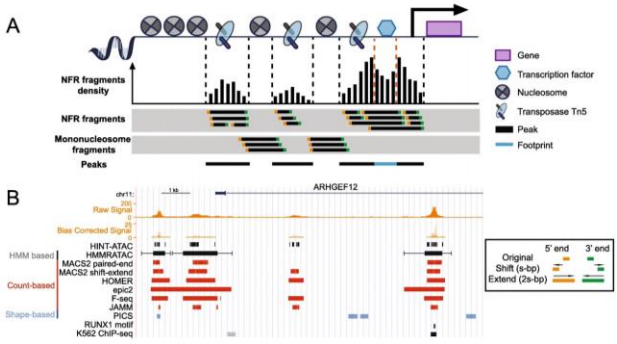




The overall data analytics pipeline for epigenetics



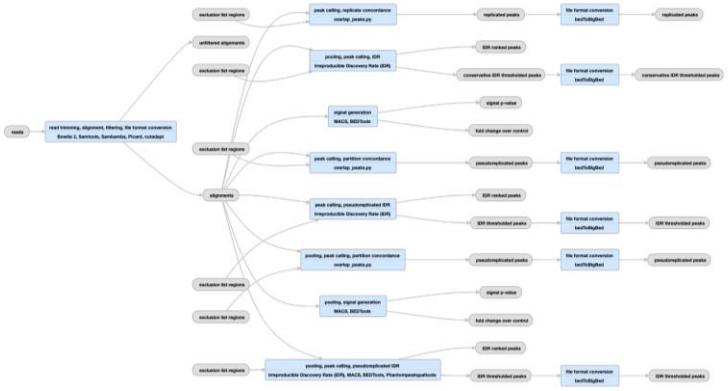
Peak calling



The output of peak calling(Browser Extensible Data (BED) format)

```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

The Entire Detailed Pipeline (ATAC-seq as an example)



Histone marks and chromatin accessibility