BMEG 3105 Fall 2024

Data analytics for personalized genomics and precision medicine

Lecturer: Yu LI (李煜) from CSE

Lecture 16: Cancer genomics overview &genomics analysis (30/10/2024)

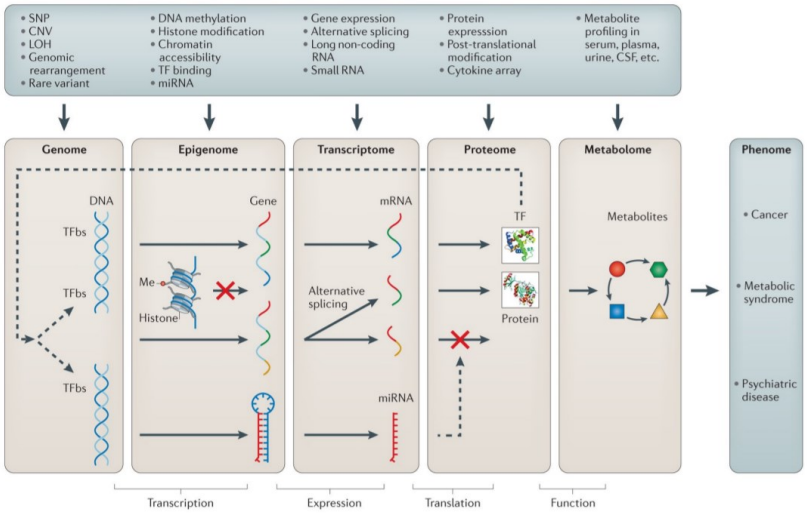
Scriber: Tom Yuet Yi (1155160583)

Cancer

- Disease with uncontrollable cell growth
- Symptom of spreading to other body parts
- Often identified as a genomic disease

➔ Study through genomics and multi-omics.

Multi-omics level: genome → epigenome → transcriptome → proteome → metabolome

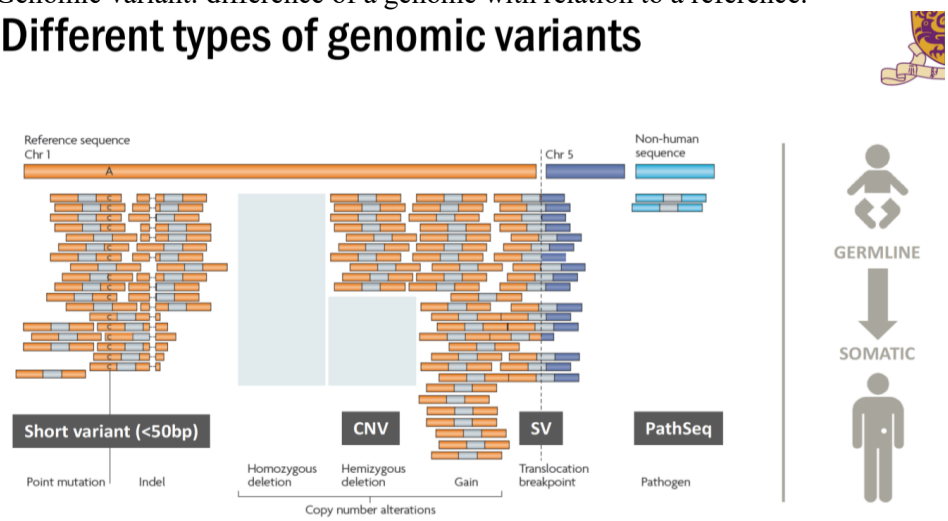


Data Analytics for Cancer Genomics

Genome: variant calling, genome-wide association studies (GWAS)

- Variant calling

  Genomic variant: difference of a genome with relation to a reference.



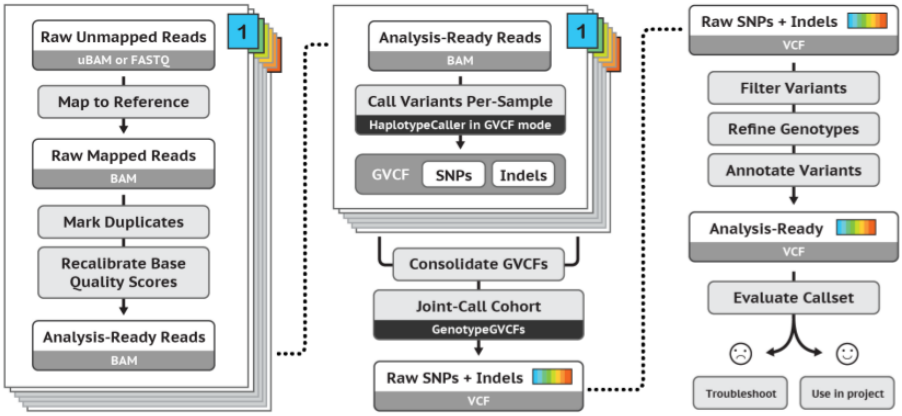
➔ Contribute to: disease risk, response to treatment, variants that contribute to disease

➔ Cancer have genetic variants at multiple levels.

Genetic variant or Errors?
Genetic variants: the actual change // Errors: artifacts that introduce to the analysis.
Errors can be found at:
- PCR artifacts
- Sequencing
- Alignment
- Variant calling
- Genotyping

-

Steps of discovering the genetic variants:
Enormous pile of short reads in terms of uBAM or FASTQ (Library preparation, Sequencing) → Reads mapped and cleaned up (Data pre-processing steps: 1) mapping, 2) marking duplicates, 3) base recalibration) → obtain the list of variants



1) Map the reads produced to reference:
input as uBAM or FASTQ, output as SAM or BAM
Alignment structure summarised by CIGAR

2) Mark duplicates to mitigate duplication artifacts (duplicates are non-independent measurements of a sequence fragment and must be removed for alleles correctly)
***Where does duplicates from?? → Library (during PCR) or optical duplicates (during sequencing)*

- Genome-Wide Association Studies (GWAS)
aim to identify associations between genetic variants and traits or diseases.
The sheer number of SNPs (e.g., 3.5 million) complicates the detection of significant associations, solution? → Bonferroni correction: Adjusted p-values are calculated to account for multiple testing, reducing the chance of false positives.

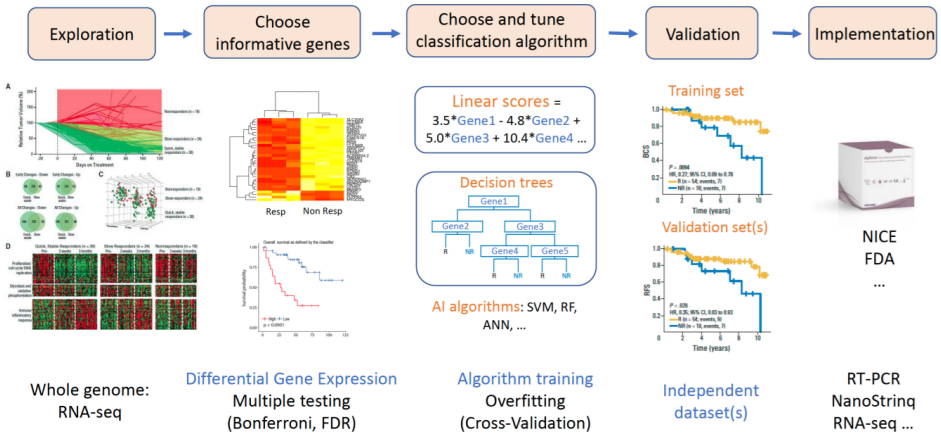
❖ **Adjusted p-value = p-value/number of tests**

❖ **Suppose we have 1 million SNPs to test**

➢ **Adjusted p-value** = $\frac{0.05}{1{,}000{,}000}$

➢ **Adjusted p-value** = $5 * 10^{-8}$

1. -Seq Data Analysis

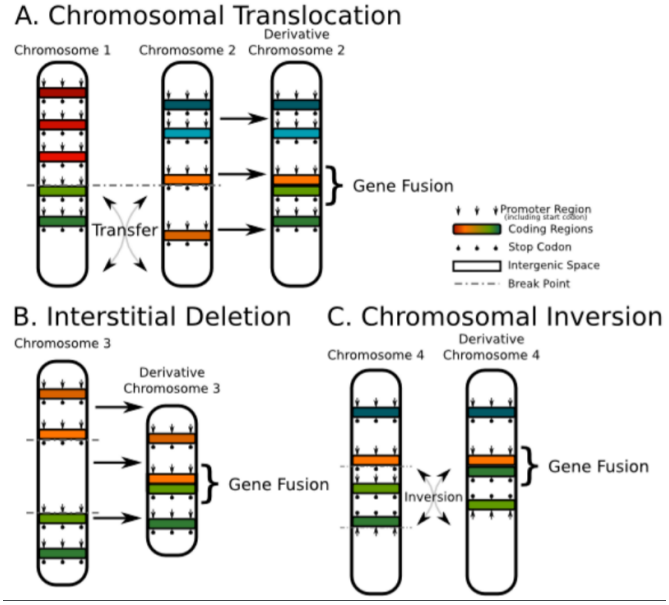


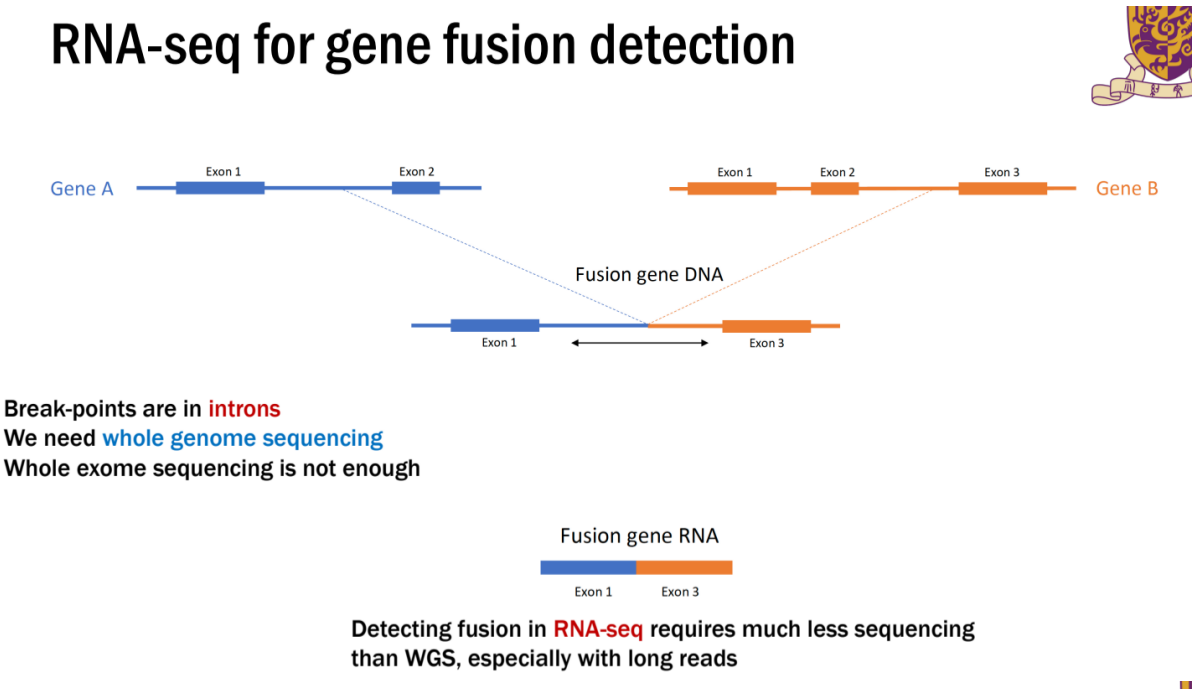
Gene Fusion – a specific kind of structural variant related to cancer

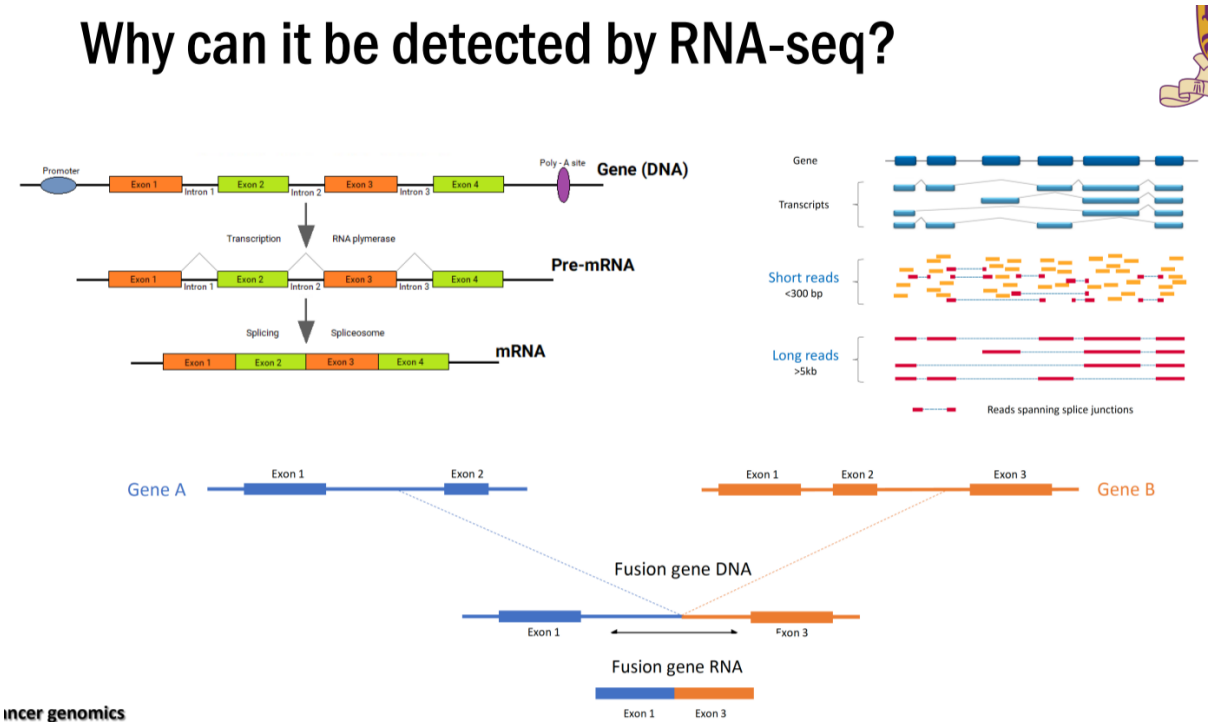
the first fusion gene was described in cancer cells

Novel gene formed by fusion of 2distinct wild type genes

in cancer: produced by somatic genome rearrangement





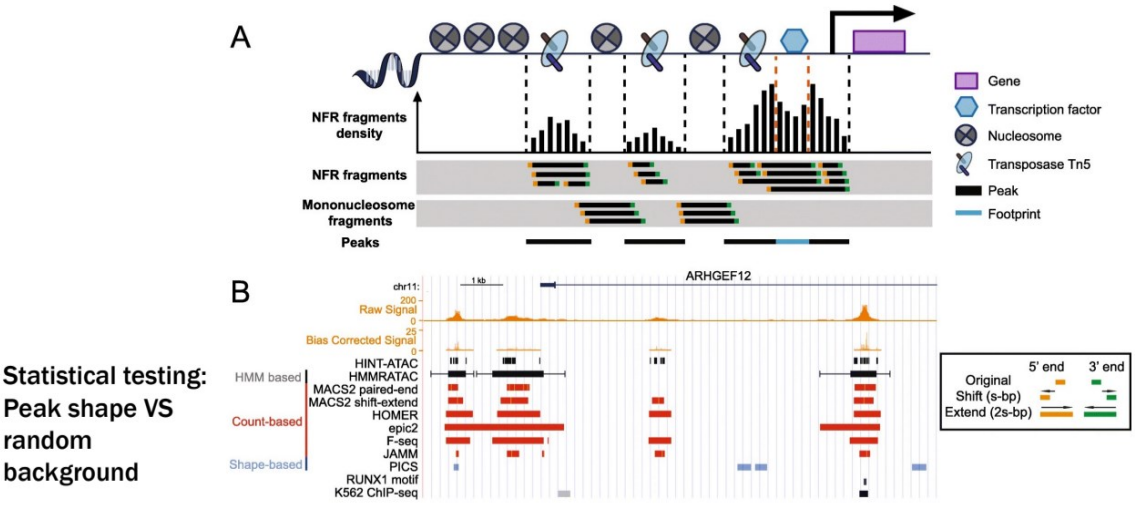


2. Epigenomics

Peak Calling:

- Peak calling is a key step in the epigenomics data analysis pipeline.
- It aims to identify regions of the genome with significantly higher signal compared to the background.
- This involves statistical testing to determine which peaks are genuine vs. random background.
- statistical testing is used to compares the peak shape to the random background signal.

- This allows distinguishing true peaks from noise.

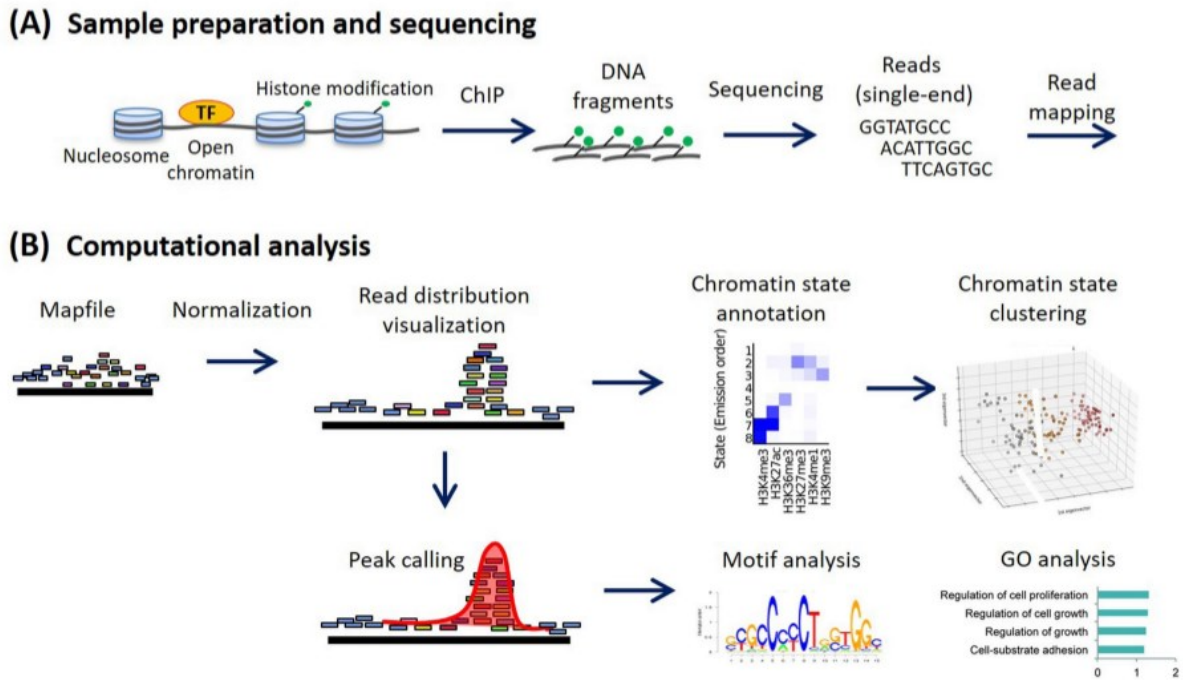


Statistical testing: Peak shape VS random background

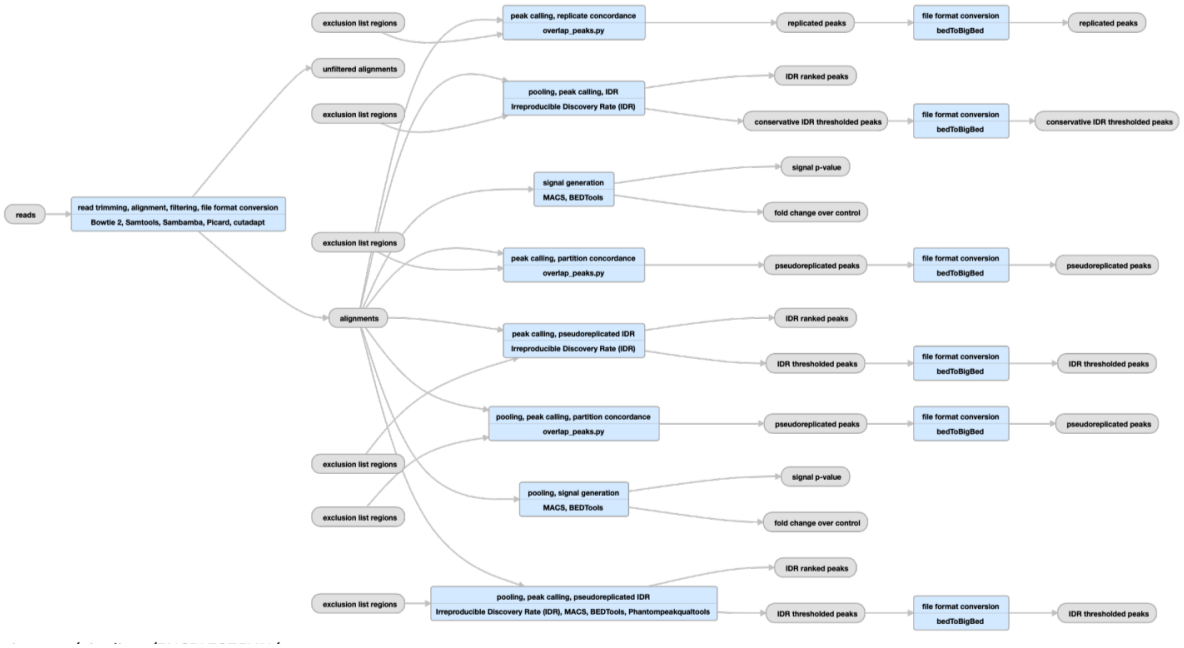
Peak Calling Output:

- The output of peak calling is typically in Browser Extensible Data (BED) format.
- This includes the chromosome, start, end, and label/annotation for each identified peak.

The Epigenomics Data Analysis Pipeline:



The Detailed Epigenomics Pipeline:


reads
read trimming, alignment, filtering, file format conversion
Bowtie 2, Samtools, Sambamba, Picard, cutadapt
unfiltered alignments
alignments
exclusion list regions
exclusion list regions
exclusion list regions
exclusion list regions
exclusion list regions
exclusion list regions
peak calling, replicate concordance
overlap_peaks.py
replicated peaks
file format conversion
bedToBigBed
replicated peaks
pooling, peak calling, IDR
Irreproducible Discovery Rate (IDR)
IDR ranked peaks
conservative IDR thresholded peaks
file format conversion
bedToBigBed
conservative IDR thresholded peaks
signal generation
MACS, BEDTools
signal p-value
fold change over control
peak calling, partition concordance
overlap_peaks.py
pseudoreplicated peaks
file format conversion
bedToBigBed
pseudoreplicated peaks
peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR)
IDR ranked peaks
IDR thresholded peaks
file format conversion
bedToBigBed
IDR thresholded peaks
pooling, peak calling, partition concordance
overlap_peaks.py
pseudoreplicated peaks
file format conversion
bedToBigBed
pseudoreplicated peaks
pooling, signal generation
MACS, BEDTools
signal p-value
fold change over control
pooling, peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR), MACS, BEDTools, Phantompeakqualtools
IDR ranked peaks
IDR thresholded peaks
file format conversion
bedToBigBed
IDR thresholded peaks