BMEG 3105: Data Analytics for Personalized Genomics and Precision Medicine

Lecture 16 – Cancer Genomics Overview

Lecturer: Professor Li Yu

Date: 28 October

TAO Xinyao
1155157215

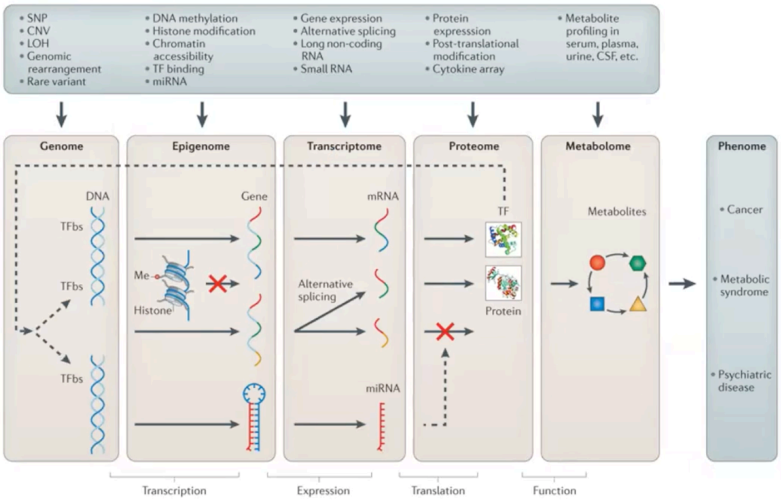
Contents

1. Review last lecture

-How to deal with overfitting?

Data; Model; Connectivity; Parameter value range; Training time.

-What is multi-omics



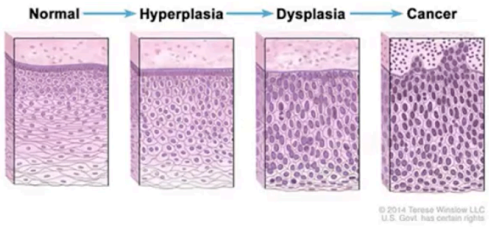
-Differential gene expression analysis

T-test; Testing association.

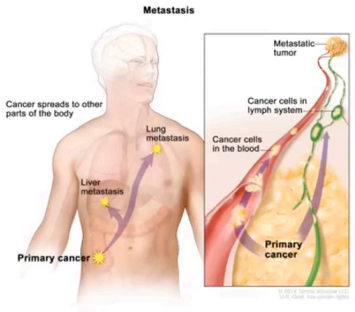
2. Cancer genomics overview

-What is cancer?

A disease ---> some of the body cells grow uncontrollably & spread to other parts of the body.



Cells grow uncontrollably



Cancer cells spread

-Why do we want to study cancer?

Cancer causes lots of death (larger than covid-19).

-How do we study cancer?

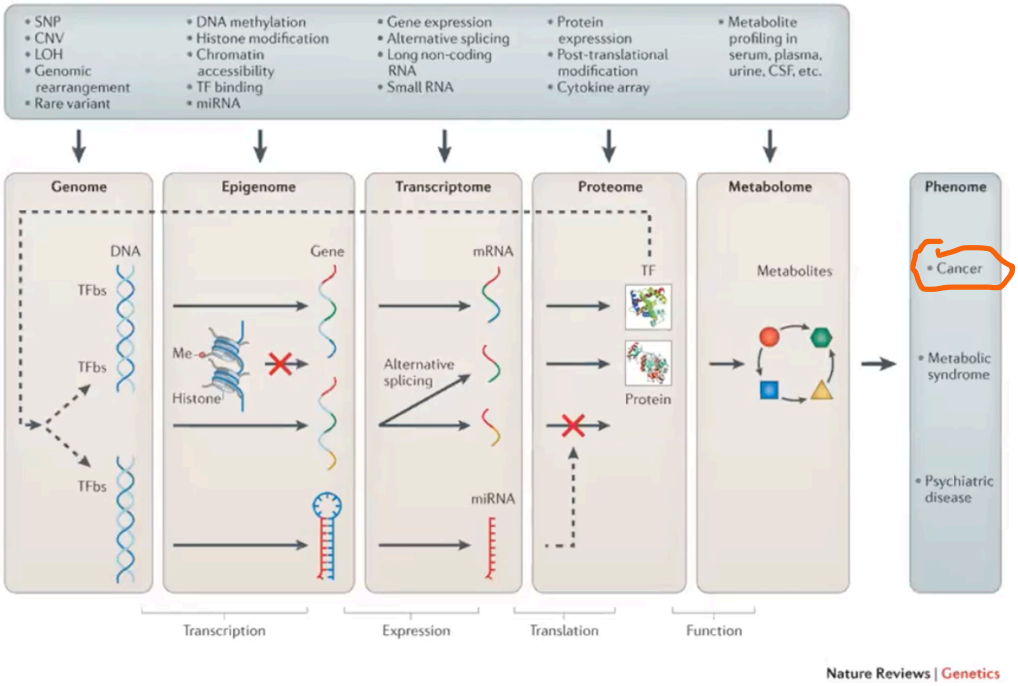
Cancer is usually believed to be a genomic disease.

--->

Use genomics/multi-omics methods to study it.

--->

Data: Genome / Epigenome/ Transcriptome/ Proteome/ Metabolome



-Data analytics for cancer genomics

a. Genome: Variant calling, genome association study

Identify mutation.

Check correlation between such mutation against cancer.

b. Epigenome: what is it, peak calling, differential peak calling

There may not be gene mutation but gene expression of the cells are different.

Gene expression regulation.

c. RNA-seq: DEG, gene fusion

3. Genome

3.1 Variant calling (very complicated nowadays)

-3.2 billion sites in the human genome.

Any two humans share 99.5% DNA.

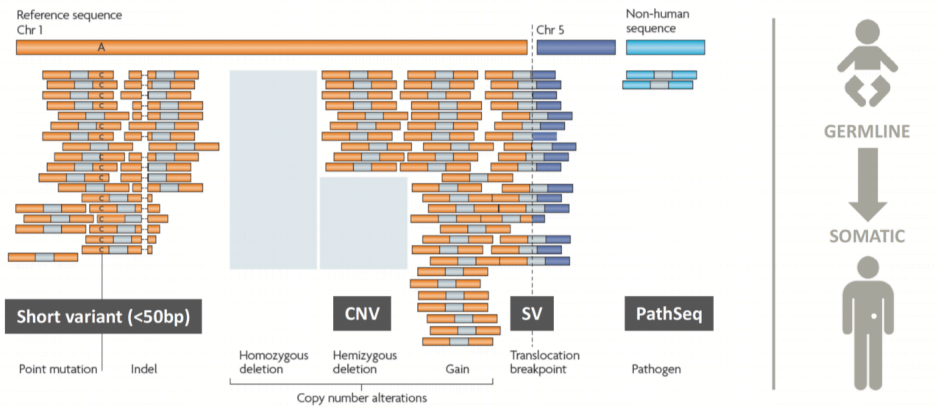
We can efficiently describe a genome with relation to a reference.

-Genetic differences among people lead to differences in disease risk and response to treatment.

-Genetic variation is used to find genes and variants that contribute to disease.

-Cancer: genetic variants at multiple levels.

3.1.1 Deferent types of genomic variants



-Short variant(<50bp)

Point mutation: one base change

Indel: insert a base

-Copy number variant: entire gene duplicated

-Structure variant

Shift

Entire gene deleted

-PathSeq

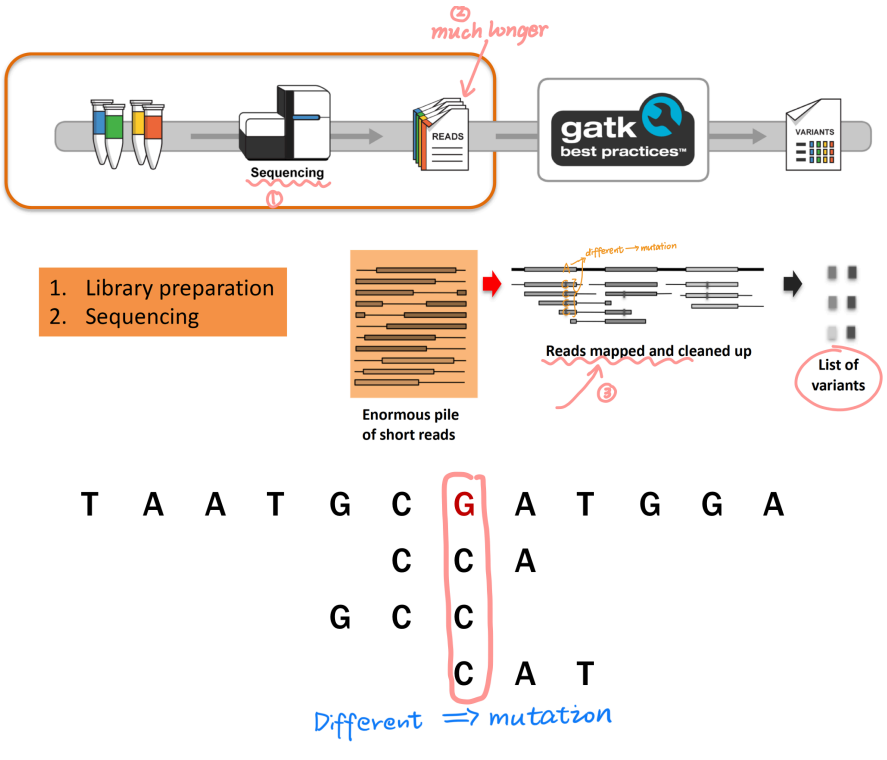
Non-human sequence inserted

-Germline vs Somatic

Germline: heritable

Somatic: not heritable

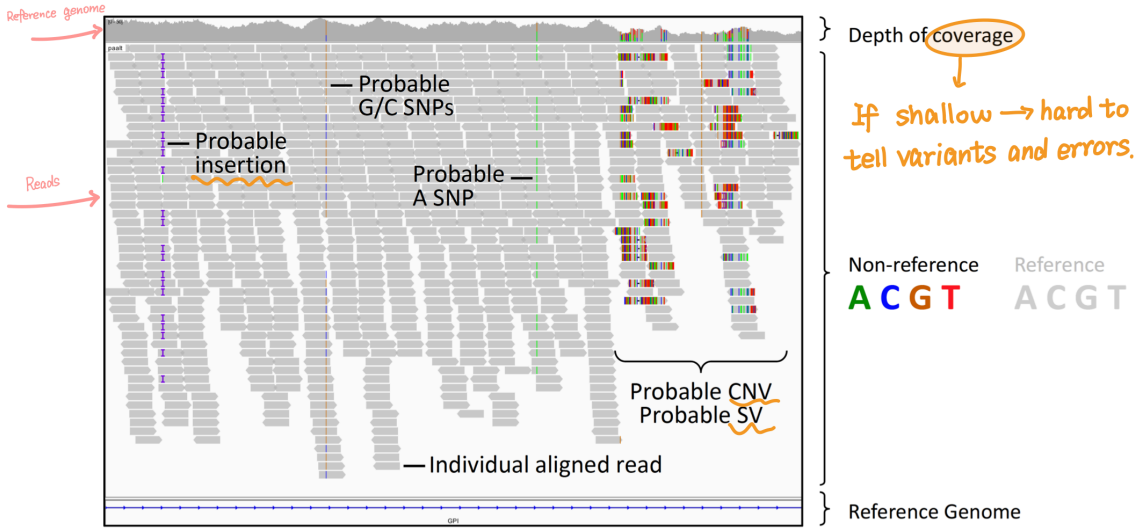
3.1.2 How to discover the genetic variants?



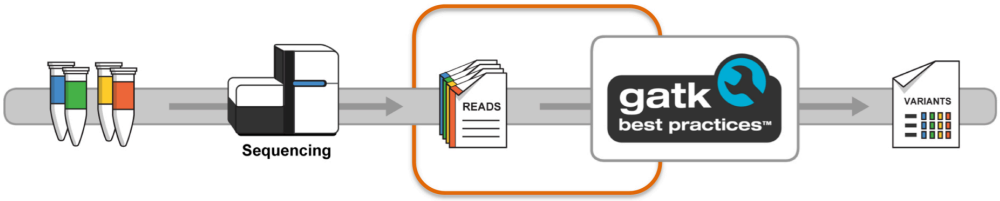
During sequencing process:Variants vs Errors
-one read: usually error; many reads: variant
-errors can creep in on various levels:

- 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)

What variants look like in a genome browser:



3.1.3 Data pre-processing step
Step1: Mapping
Step2: Marking duplicates
Step3: Base recalibration



1. Mapping
2. Marking duplicates
3. Base recalibration