

Data analytics for personalized genomics and precision medicine

Lecture 17: Genomics data analysis

Lecturer: Yu LI (李煜) from CSE

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

Wednesday, 2 November 2022

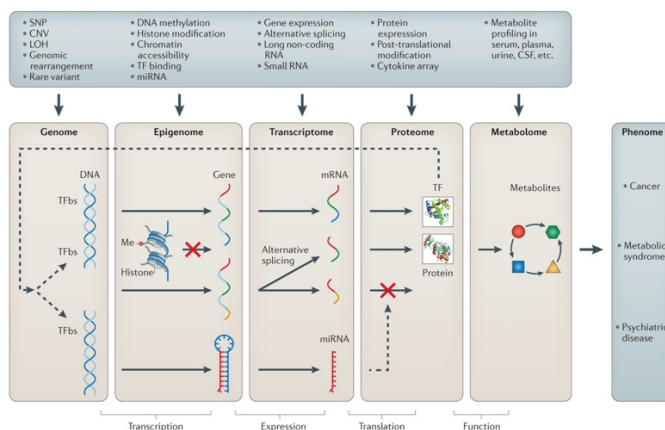
❖ Outline of the Lecture:

- About previous lecture
- Overview of this lecture
- Genome
- RNA-Seq

1. About previous lecture

❖ More about cancer

- Definition of cancer:
 - Disease in which some of the body's cells grow uncontrollably and spread to other parts of the body
- Study cancer at multiple level:



- Genetic variants
 - Genome
 - Gene fusion (RNA-seq)
- Abnormal gene expression

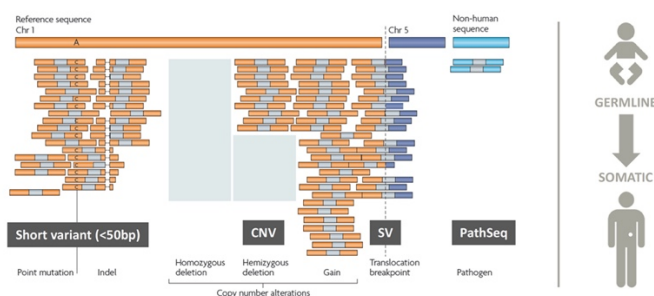
- Genome (genetic information)
 - ◆ Can be differed even when the DNA is not different
- Epigenome (environment)
- Transcriptome (direct measurement)

2. Overview about this lecture

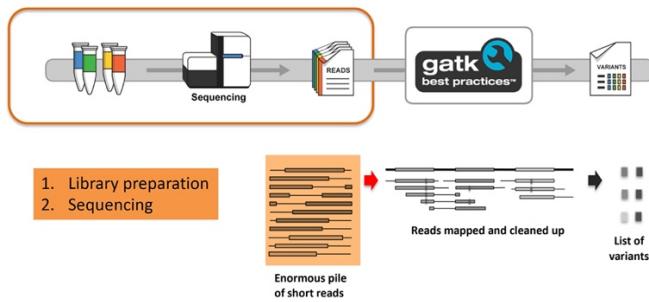
- ❖ Data analytics for cancer genomics
 - Genome: variant calling, genome association study
 - Epigenome: what is it, peak calling, differential peak calling
 - RNA-seq: DEG, gene fusion

3. Genome

- ❖ Variant calling
 - Reason that variants are important
 - 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
 - Different types of genomic variants



- How to discover genetic variants (1st part)

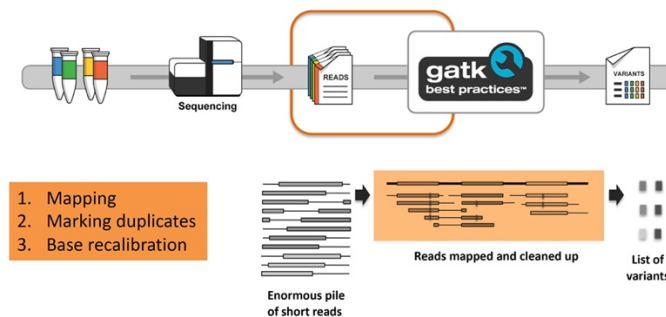


- Library preparation and Sequencing

- Variants VS errors

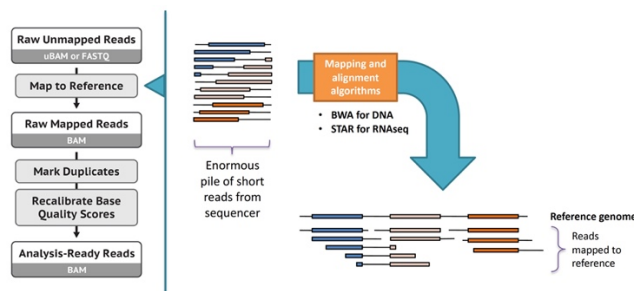
- Important to distinguish between actual variation (real change) and errors (artifacts)
- Errors creep in on various levels
 - PCR artifacts (amplification of errors)
 - Sequencing (errors in base calling)
 - Alignment (misalignment, mis-gapped alignments)
 - Variant calling (low depth of coverage, few samples)
 - Genotyping (poor annotation)

- Data pre-processing step (2nd part)

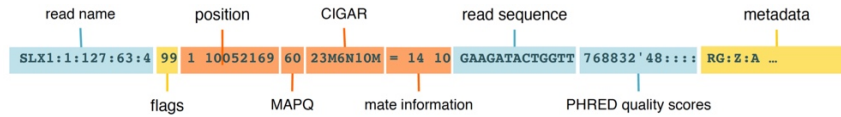


- Mapping, Marking duplicates, Base recalibration

- Step 1: Map the reads produced by the sequencer to the reference



- Input format (FASTQ)



- Added mapping info summarizes **position, quality, and structure** for each read
- Mate information points to the read from the other end of the molecule (other in a pair)

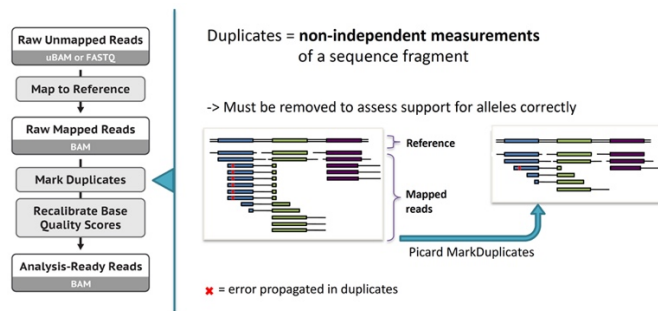
• **CIGAR (Concise Idiosyncratic Gapped Alignment Report)**

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set '*' if unavailable):

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

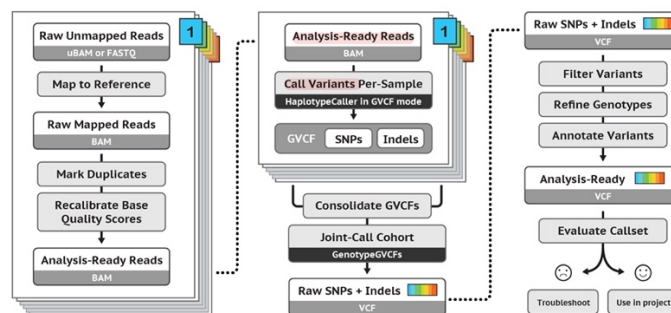
- "Consumes query" and "consumes reference" indicate whether the CIGAR operation causes the alignment to step along the query sequence and the reference sequence respectively.
- H can only be present as the first and/or last operation.
- S may only have H operations between them and the ends of the CIGAR string.
- For mRNA-to-genome alignment, an N operation represents an intron. For other types of alignments, the interpretation of N is not defined.
- Sum of lengths of the M/I/S/=/X operations shall equal the length of SEQ.

▪ **Step 2: Mark duplicates to mitigate duplication artifacts**



- Error may be due to noise/ error/ duplicate

▪ **Variant calling**

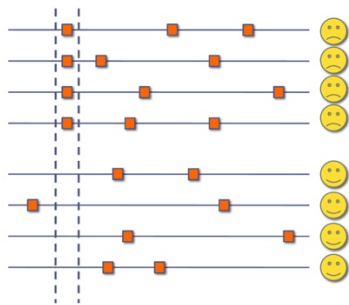


- Joint analysis empowers discovery
- Single genome in isolation → almost never useful
- Family or population data → add valuable information
 - Rarity of variants

- De novo mutations
- Ethnic background

❖ GWAS (Genome-wide association studies)

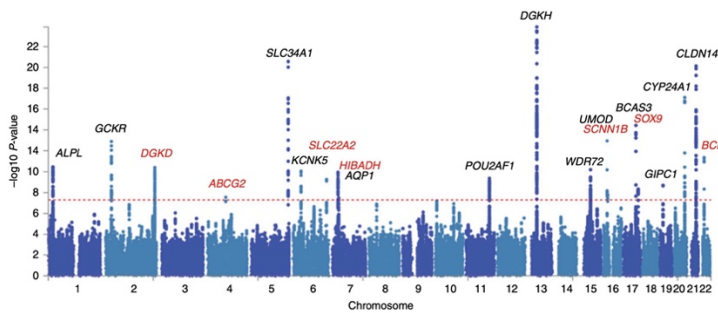
- Trying to determine whether specific variant(s) in many individuals can be associated with a trait (disease)



Spot the variant that is common amongst all affected but absent in all unaffected

The ideal case (for some rare Mendelian diseases)

- In reality: 3.5 million SNPs

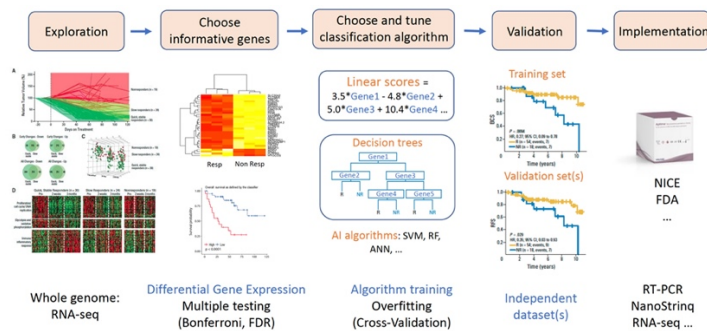


- Bonferroni correction

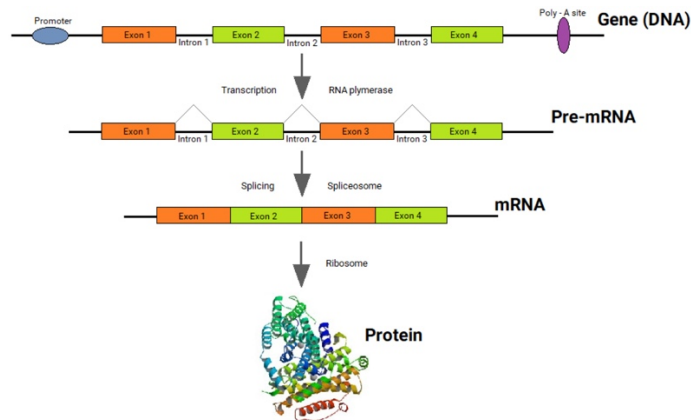
- Adjusted p-value = p-value / number of tests

4. RNA-seq

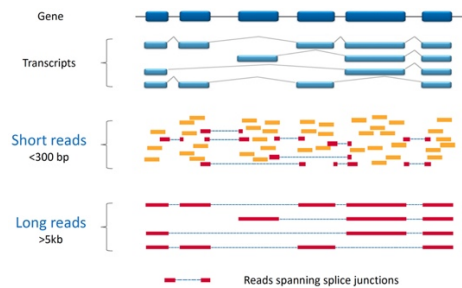
❖ RNA-seq data analysis



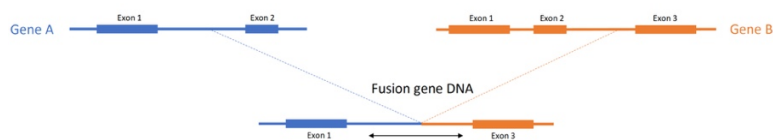
❖ Transcription, splicing and translation of a eukaryotic gene



❖ Mapping spanning splice junctions



➤ Gene fusion



- The first gene was described in cancer cells in early 1980s
- Novel gene formed by fusion of two distinct wild type genes
- In cancer: produced by somatic genome rearrangements

❖ RNA-seq for gene fusion detection

- Break-points are in introns
- Need whole genome sequencing, whole exome sequencing is not enough
- Detecting fusion in RNA-seq requires much less sequencing than WGS
- Can be detected by RNA-seq