BMEG3105 Data analytics for personalized genomics and precision medicine

Week 6 Lecture 10

Cancer Genomics

By Cheung Ho Lun 1155174348 11/10/2024

1. About last lecture

Overfitting and underfitting

T-test (Purpose, general idea, formula)

Gene enrichment analysis (testing association)

Contingency tables

	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213

Fisher's exact test

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

2. Cancer genomics overview

Cancer: It is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body.

How to study?

➔ As cancer is believed to be a genomic disease, we can use multi-omics.

Genome -> Epigenome -> Transcriptome -> Proteome -> Metabolome

Genome: **variant calling**, genome association study

Epigenome: peak calling, differential peak calling

Exercise (UReply): Which would not be affected by epigenomic modification?

A. DNA sequence

B. Gene expression

C. The transcription process

D. Disease and phenotype

3. Variants

*Why do we care about **variants**?*

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

Different types of genomic variants:

Short variant (point mutation, insertion, deletion) -> Copy number variant ->

Structure variant -> Path sequence

Steps to discover the genetic variants:

I. Library preparation

II. Sequencing

(Recall the sequencing mapping from previous lecture)

Important: Difference between variants (real change) and errors (artifacts)

Error can creep in various level:

PCR artifacts (amplification of errors)

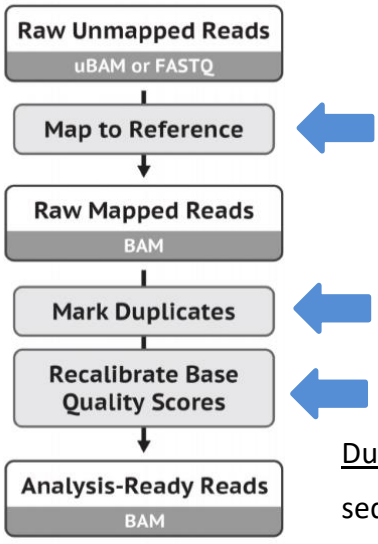
Sequencing (errors in base calling)

Alignment (misalignment, mis-gapped alignment)

Variant calling (low depth of coverage, few samples)

Genotyping (poor annotation)

4. Data pre-processing step



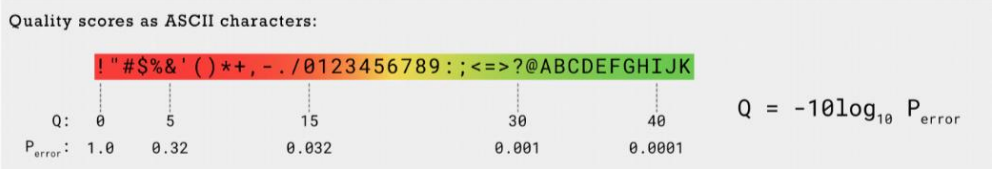
1. Map the reads produced by the sequence to the reference

2. Mark duplicates to mitigate <u>duplication</u> artifacts

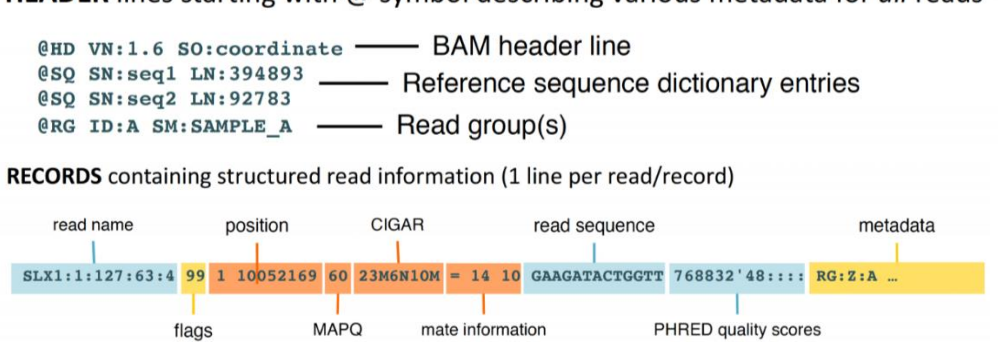
3. Base recalibration

<u>Duplicates</u> = **non-independent measurements** of a sequence fragment. It must be removed to assess support for alleles correctly

<u>Input format: FASTQ **(How to read?)**</u>



Output format: Sequence/ Binary alignment Map (SAM/BAM)



- The information in the head belongs to mate information.
- MAPQ is about quality.
- Added mapping information summaries position, quality, and structure for each read.
- Mate information points to the read from the other end of the molecule.

Important: 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
```

How to read?

➔ 3M: 3 matches, 1D: 1 deletion, 2M: 2 matches, 1I: 1 insertion, 1M: 1 match

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

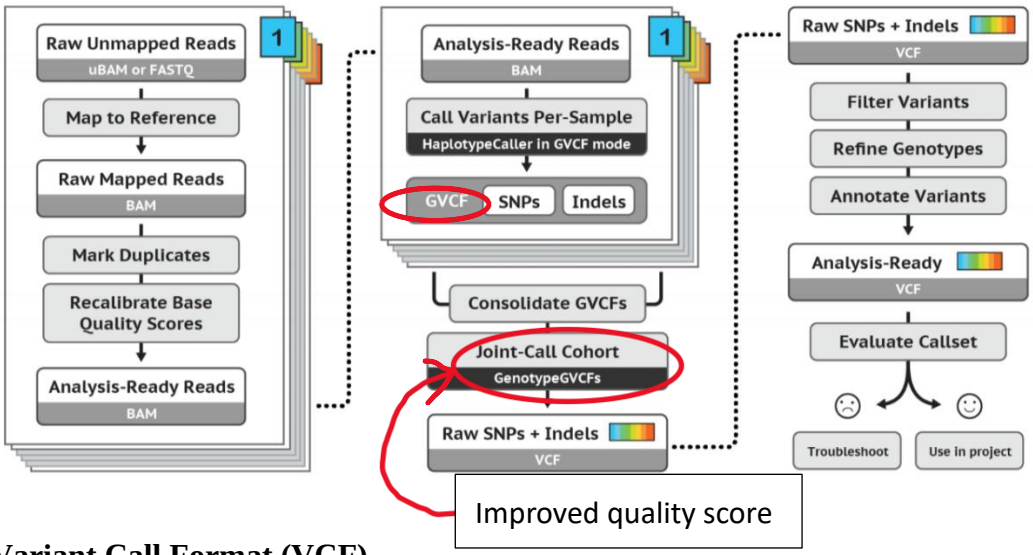
H can only be present as the first and/or last operation.

S may only have H operations between them and the ends of the CIGAR string.

Where does the duplication come from?

➔ Library duplicates cause by PCR/ Optical duplicates occurred in sequencing

5. Variant calling



Variant Call Format (VCF)

```
HEADER
##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
RECORDS
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
```

Joint analysis empowers discovery

Single genome in isolation: almost never useful

Family or population data add value information

➔ Rarity of variants

➔ *De novo* mutations

➔ Ethnic background

Note: After multi-sample VCF, the quality score will be higher because it combines information from samples as illustrated above.

Exercise (UReply): Which will affect the variant calling step?

A. Fastq file format

B. Pre-processing quality control

C. GWAS corrected p-value

D. VCF file format

6. Genome-wide association studies (GWAS)

➔ Trying to determine whether specific variants in many individuals can be associated with a trait (disease)

➔ Bonferroni correction

Given a 3.5 million SNPs, it we use p = 0.05. Then, the probability of the result being correct would be $(1 - 0.05)^{3.5\ millions}$, which is almost zero. **This means at least one testing result is wrong.**

Therefore, we need to adjust p-value

p-value = p-value/ number of tests

e.g. we have 1 million SNPs to test

Adjusted p-value = $\frac{0.05}{1000000}$ = 5 * 10^{-8}