BMEG 3105 Fall 2024

Data analytics for personalized genomics and precision medicine

Lecture 15 Scribing

Lecturer: Yu LI (李煜) from CSE

Wednesday 30 October 2024

Lecture agenda:

- Recap of last lecture
- Cancer genomics overview
- Genome – Variant calling & GWAS

Expected outcomes:

- When doing pipeline, able to know each step and the expecting file
- Able to utilize the tool practically
- Can troubleshoot and know what to input into a specific step
- Understand the reasons for each step
- The ability to read the records in different files
- Different factors which affect the quality of the mapping and the variant calling

Feedback and comments from last lecture:

- Generally positive feedback without further requirement

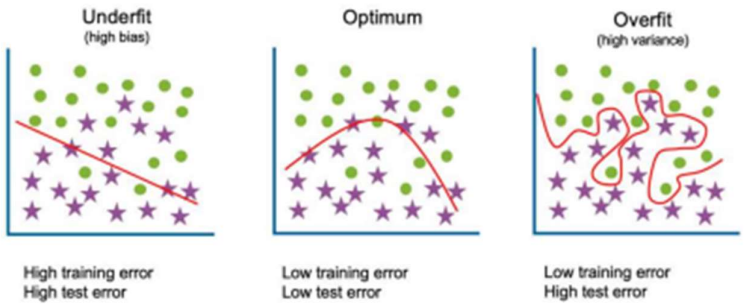
Recap:

Underfitting & Overfitting:

- Underfitting:
  - Definition: The relationship among different variables within the image is more complicated than simple linear combination
  - This leads to the model capacity is not enough
- Overfitting:
  - Definition:
    Statistically: The production of an analysis that corresponds too closely or exactly to a particular set of data, and may lead to failure to fit additional data or predict future observations reliably

    Machine learning: The method is too complex to the problem, which may perform well on the training dataset but not on the testing dataset.

*In practice, performing in testing dataset is more important.



Multi-omics:

- Definition:
  The data sets of different omics groups are combined during computational analysis.
- Core techniques:
  - Sequence alignment and comparison
  - Dimension reduction and visualization
  - Clustering and classification

Differential Gene Expression Analysis:

- Purpose:
  Using statistical analysis to discover quantitative changes in gene expression levels between experimental groups.
  [e.g. Whether the gene expression difference is significant, other than due to natural random variation.]

- Method of analysis:
  - T-test
    Purpose: Discover the significance difference between two data sets.
    Details:
    a. Calculate a test statistic based on the mean and variance of the data
    b. Test statistics follow a Student's t-distribution
    c. P-values: the probability that the result from the data occurred by chance (the smaller the p-values, the more confident we are)

    Different kinds of T-test:

    a. Various formulas to calculate t-values
    b. Various formulas to translate t-value to p-values

    *If p-values are smaller than 0.05, we define the two sets of data are different.

  - Gene enrichment analysis
    a. Testing association
    b. Contingency tables

Pathway VS Gene mutation/expression

	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

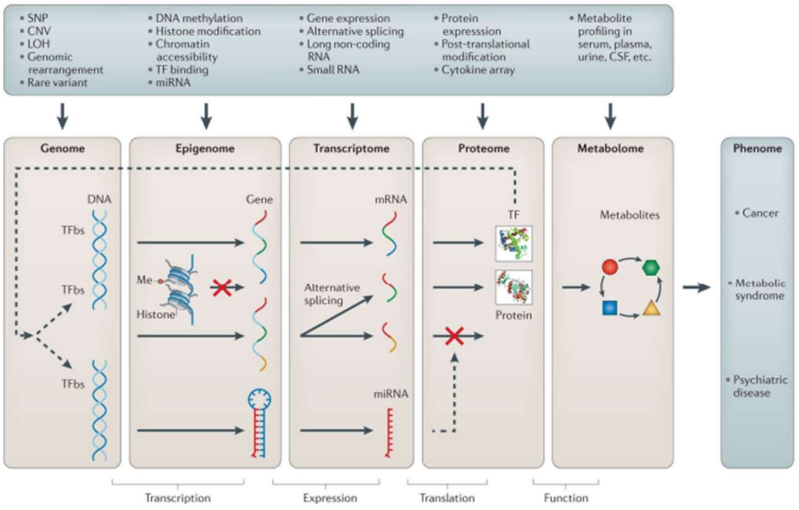
*If they are related – a, d should be large while b, c should be small
*To define that they are related, we need further confirmation/data:
1. quantitative measure
2. A standard procedure
3. Statistical test for association
4. Fisher's exact test
> definition: Fisher's exact test is statistical significance test used in the analysis of contingency tables.
>P-values can be calculated directly from the table

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

Lecture:

Cancer Genomics Overview:

- Definition: body cells are in the stage of uncontrollably continuous deviation and spread to other parts of the body.
- Significance:
  - Many different type of cancer
  - One of the major leads to death
- Method of studying cancer:
  - Defined as a genomic disease, therefore, use genomics/multi-omics methods
  - Including genome/epigenome/transcriptome/proteome/metabolome

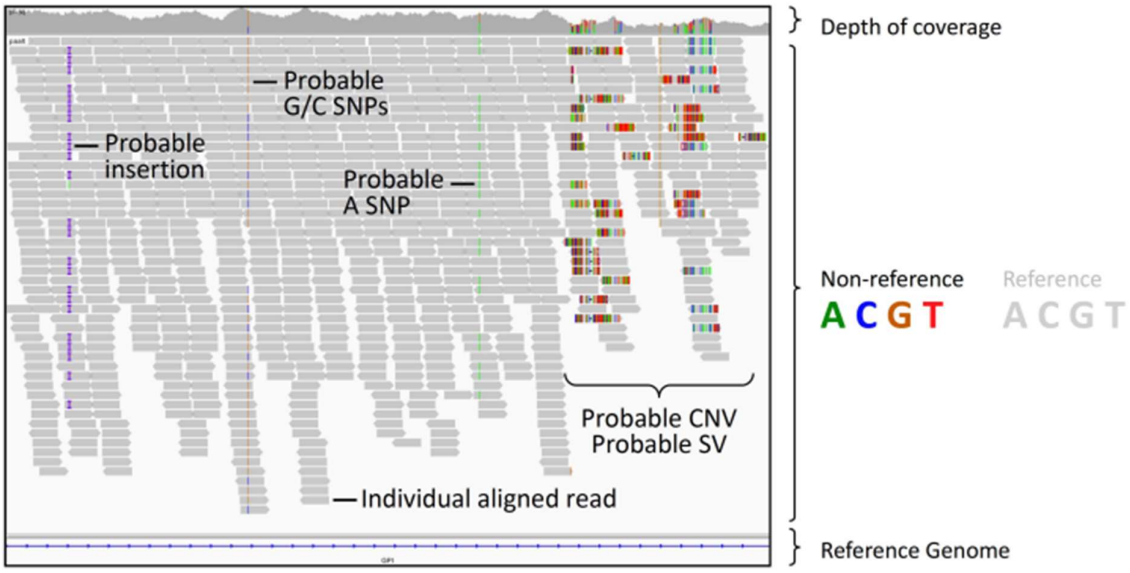


- Data analytics for cancer genomics:
  - Genome: variant calling, genome association study
  - RNA-seq: DEG, gene fusion
  - Epigenome: what is it, peak calling, differential peak calling

Genome:

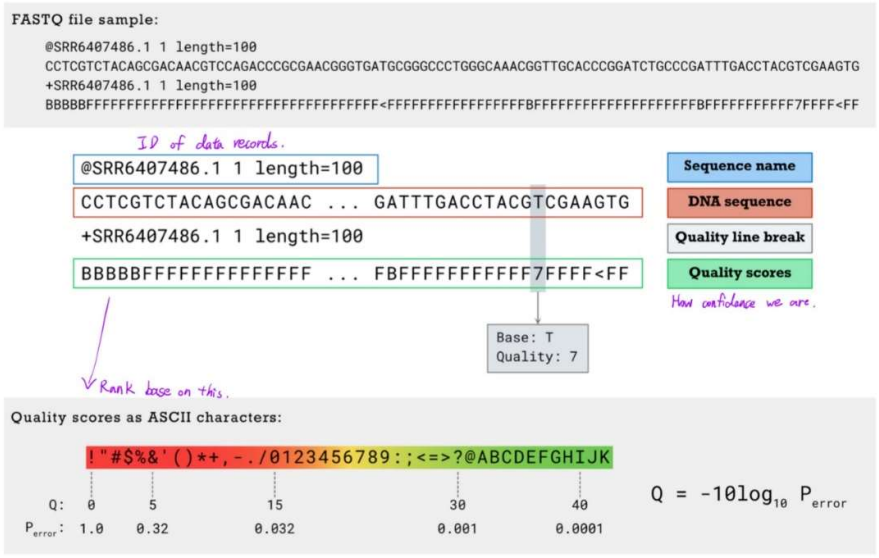
- Variant calling
  - Significance:
    a. Describing a genome with relation to a reference
    b. Genetic differences among people lead to differences in disease risk and response to treatment
    c. Genetic variation is utilize to identify the genes and variants that contribute to disease
    d. Specifically, the genetic variants in cancer are at multiple levels (can be signal nucleotide mutation or due to gene expression)
  - Types of genomic variants
    a. Short variant - point mutation & deletion/addition
    b. Copy number variation (CNV) – homozygous deletion, hemizygous deletion, gain
    c. Structural variants (SV) – translocation breakpoint
    d. Pathogen (PathSeq) – non-human sequence which may from virous and etc.
- Steps of genetic variants discovery
  - Library preparation
  - Sequencing
  - Base recalibration (BQSR) - A process in which we apply machine learning to model these errors empirically and adjust the quality scores accordingly
    *To distinguish the variants and errors:
    - Errors can creep in on various levels:
    1. PCR artifacts (amplification of errors)
    2. Sequencing (errors in base calling)
    3. Alignment (misalignment, mis-gapped alignments)
    4. Variant calling (low depth of coverage, few samples)
    5. Genotyping (poor annotation)

e.g. Variant example:



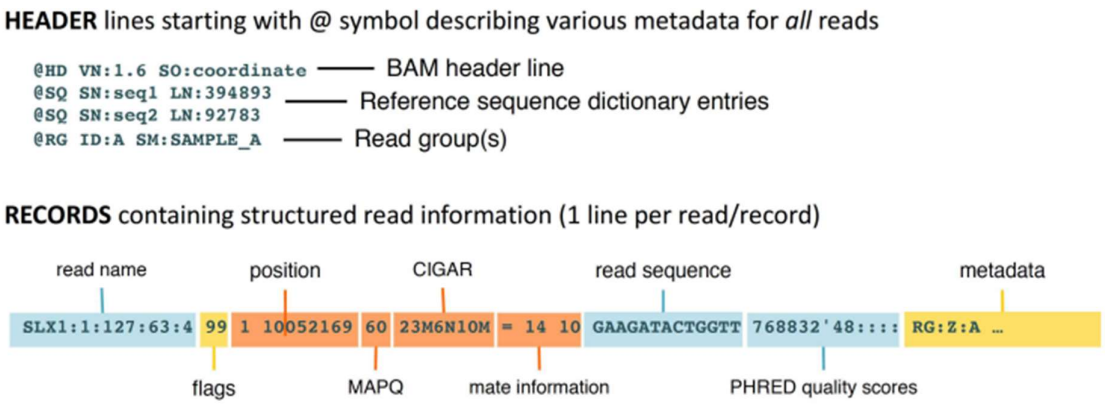
* Depth of coverage: for a spefic site, the amount of reads that are map to the region.

- Data pre-processing step:
  - Step 1: Map the reads produced by the sequencer to the reference
    Input format – FASTQ
    How to read FASTQ:



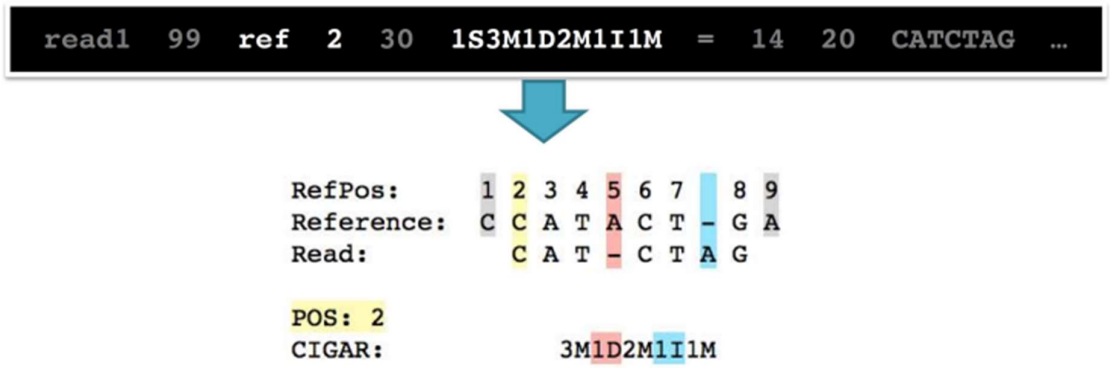
Output format: Sequence/Binary Alignment Map (SAM/BAM)

How to read SAM/BAM:



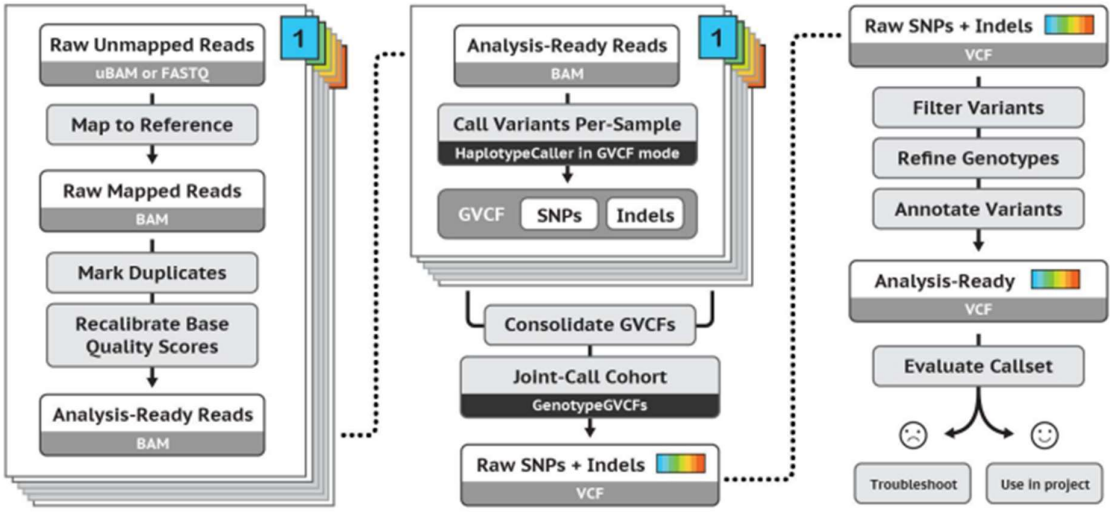
CIGAR stads for Concise Idiosyncratic Gapped Alignment Report, which summarizse alignment structure.

How to read CIGAR:



- Step 2: Mark duplicates to mitigate duplication artifacts
  Duplicates may cause over confidence, therefore, should be removed to assess support for data correctly.
  The duplicates may be come from PCR (library duplicates) or optical duplicates which occur during sequencing.

- Variant calling



  - How to read 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
```

  - The per-sample GVCFs combine to finalize in multi-sample VCF to form a joint analysis – able to empower discovery
- Further downstream analysis
  - Genome-wide association studies (GWAS)

  Determine if specific variants in individuals are associated with traits (diseases)
  - P-value < 0.05 is not useful in practical situations, therefore, Bonferroni correction has been applied.
  - Adjusted p-value = p-value/number of tests

    [e.g. Suppose there are 1 million SNPs to test: Adjusted p-value= 0.05/1,000,000.]

Potential Project – 4, 5, 6

- Genetic variant calling pipeline
- Epigenetic data processing pipeline
- Gene fusion detection pipeline

Next lecture topic:

- RNA-seq – Gene fusion: structural variant
- Epigenome – Peak calling

Supporting Links:

- Statistical testing in Python: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
- Biopython: https://biopython.org/
- Post-lecture survey: https://forms.gle/a6rjUPxAVEGXN7Cu9

Resource and related uncovered topics:

- Data distribution & Multiple testing correction: https://www.ebi.ac.uk/training/materials/cancer-genomics-materials/
- How does cancer develop & cancer types: https://www.cancer.gov/about-cancer/understanding/what-is-cancer
- GATK workshop slides: https://drive.google.com/drive/folders/1y7q0gJ-ohNDhKG85UTRTwW1Jkq4HJ5M3
- GATK workshop video: https://www.youtube.com/watch?v=sM9cQPWwvn4
- GATK workshop: https://www.youtube.com/watch?v=xw419NKqMqw
- Epigenetics: https://www.youtube.com/watch?v=IAu44BkOaSs