BMEG3105: Data Analytics for Personalized Genomics and Precision Medicine
Lecture 14: Multi-Omics & cancer genomics overview
Autumn 2024
Lecturer: LI Yu
Scribe: 1155191391

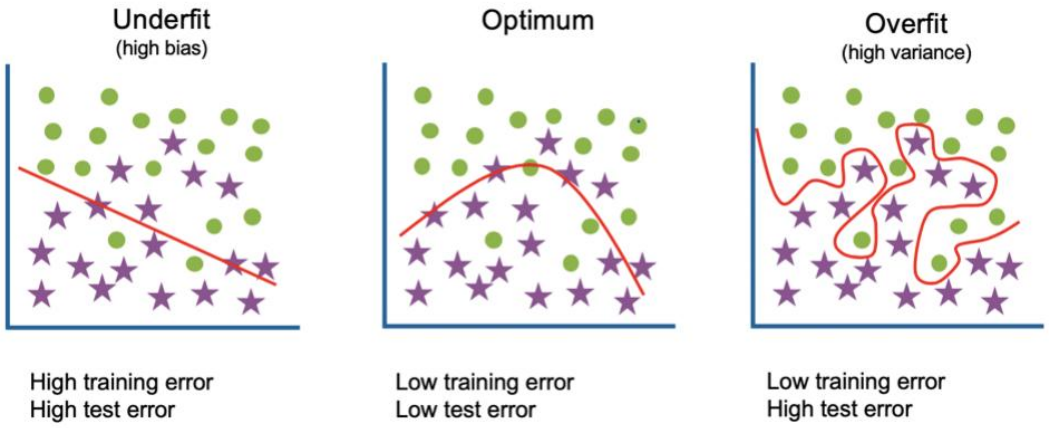
Outline of Lecture 14:

- Model underfitting and overfitting (advanced topic) (Slide: 3-13)
- Multi-omics overview (Slide: 14-27)
- Statistical testing (Slide: 28-42)
- Cancer genomics overview (Slide: 43-58)

I. Model underfitting and overfitting (advanced topic) (Slide: 3-13)

1. Underfitting:

- The model capacity is not enough
- In practice, we need to overfit the data first



Underfit, Optimum and Overfit comparison

- How to make models with more capacity?
    1. Increase the number of nodes
    2. Increase the number of layers
    3. Add non-linear function
    4. Fully-connected layers:
        - A general function approximator
        - We can approximate any function (relation) if we have enough nodes and layers
        - Universal approximation theorem

2. Overfitting:

- What is Overfitting?
    1. Statistically: the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit

additional data or predict future observations reliably

2. Machine learning: the method is more complex than the problem, such that it can perform well on the training dataset but does not perform well on the testing dataset

- How to evaluate model and detect overfitting?
    1. Train-validation-test split
    2. Cross-validation

- How to deal with overfitting?
    1. Data: Too little, not reflect the true distribution
    2. Model: Too large, too many useless parameters
    3. Connectivity: Too strong, co-adaptation
    4. Parameter value range: Too large, model too flexible
    5. Training time: Too long, tend to overfitting

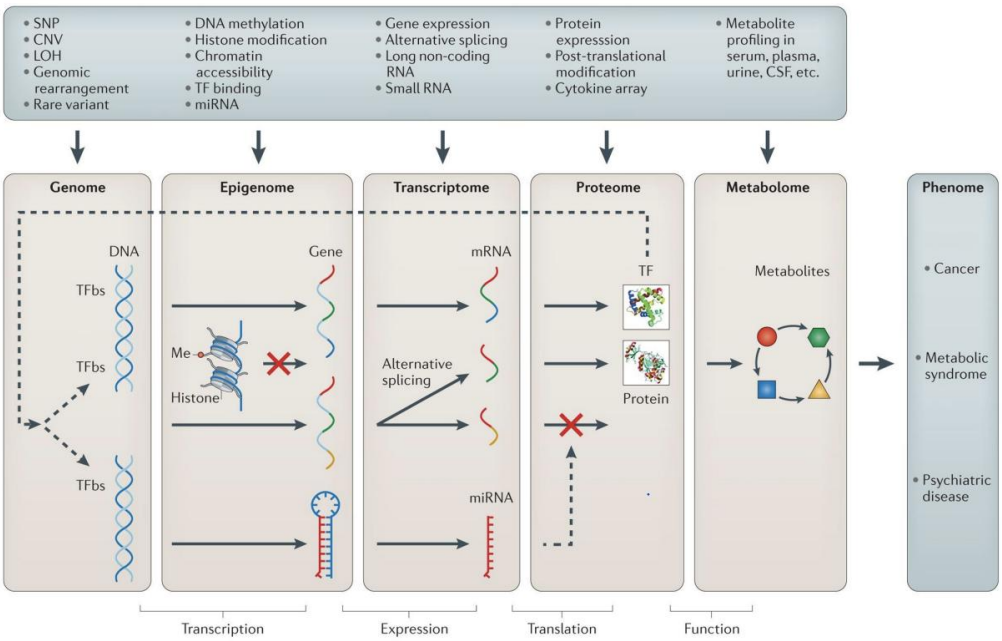
II. Multi-omics overview (Slide: 14-27)

1. What is omics?

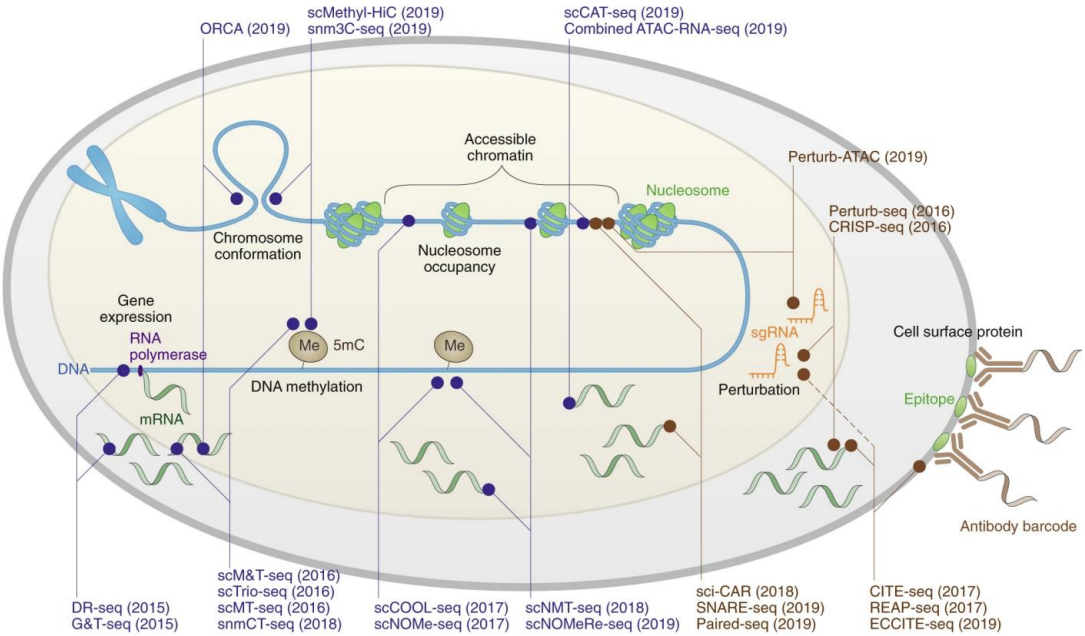
- Omics aims at the collective characterization and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism or organisms
- Study biological entities in large scale

2. What is multi-omics?

- Multi-omics

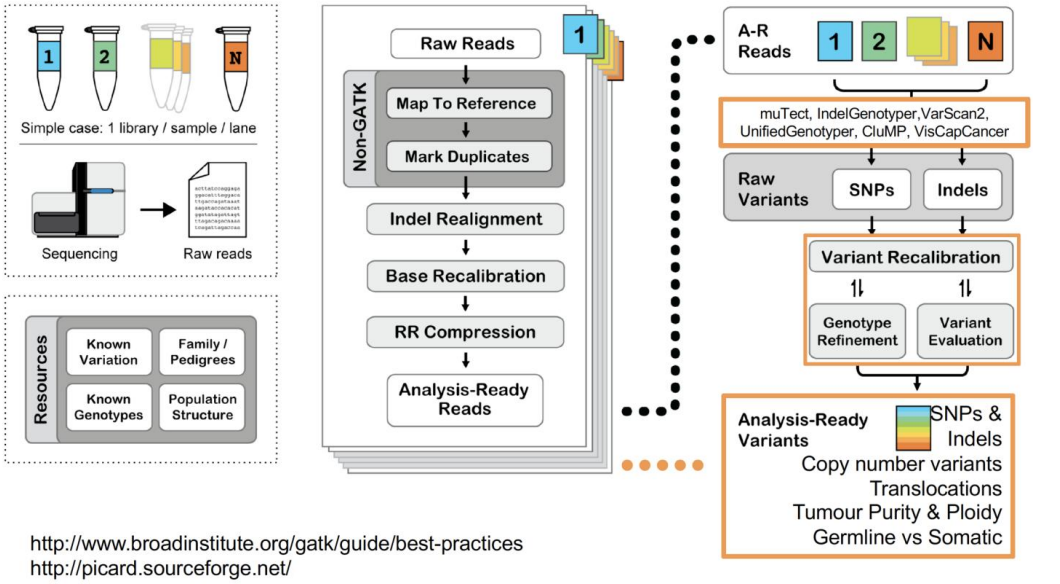


- Single-cell Multi-omics



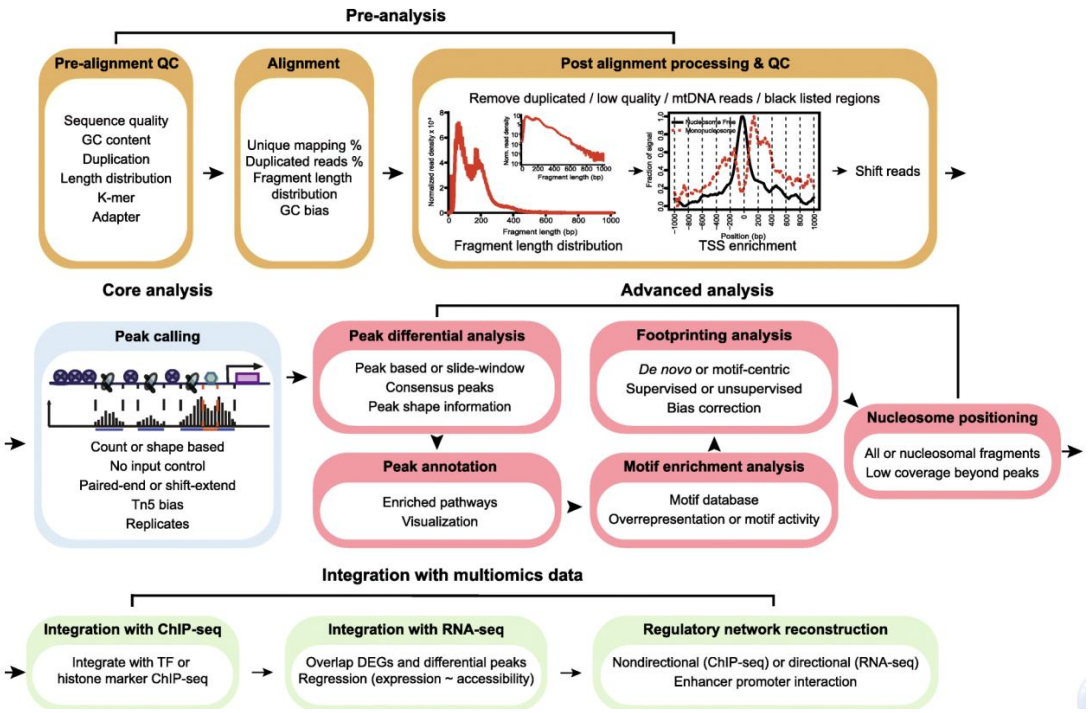
3. Pipelines:

- Genome pipeline

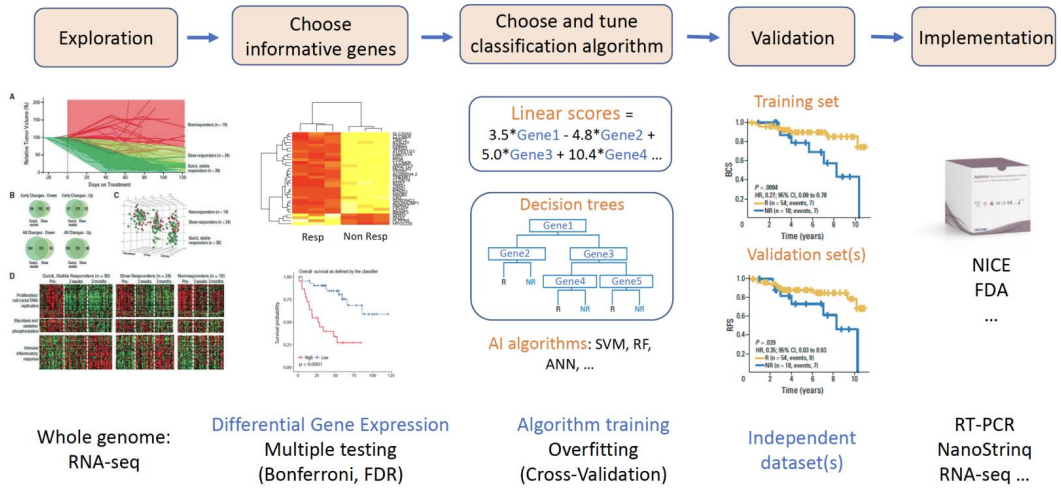


http://www.broadinstitute.org/gatk/guide/best-practices
http://picard.sourceforge.net/

- Epigenome pipeline



- Transcriptome pipeline



4. Core techniques

- Sequence alignment and comparison
- Dimension reduction and visualization
- Clustering and classification
- Statistical testing；
    1. Variant calling
    2. Peak calling
    3. Peak differential analysis
    4. Differential gene expression analysis
    5. Motif enrichment analysis
    6. GO enrichment analysis

7. KEGG enrichment analysis
8. Genome-wide association study (GWAS)
9. ...

III. Statistical testing (Slide: 28-42)

1. Differential gene expression analysis

- Statistical analysis to discover quantitative changes in expression levels between experimental groups
- For a given gene, whether the gene expression difference is significant, other than due to natural random variation

2. T-test

- A kind of standard statistical test procedure
- The purpose of t-test: Is there a significant difference between two sets of data?
- General idea:
    1. Calculate a test statistic based on the mean and variance of the data
    2. Test statistic follows a Student’ s t-distribution
    3. P-value: the probability that the result from the data occurred by chance:
        - Along with test statistic, t-value
        - The smaller p-value is, the more confident we are

- How to do T-test:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Usually, the p-value should be smaller than 0.05

- Different kinds of t-test
    - Paired VS unpaired: For paired one, we cannotshuffle the values
    - One-tailed test VS two-tailed test
        - Two-tailed test: different or the same
        - One-tailed test: greater, larger, smaller, at least

- For different t-test
    - The formula to calculate t-value can be different
    - The formula to translate t-value to p-value can be different
    - But the t-test procedure is the same
    - Eventually, we will say the two sets of numbers are different if p-value is

smaller than 0.05

3. Fisher's exact test

- Fisher's exact test is a statistical significance test used in the analysis of contingency tables

- Why is it called exact test?
  - P-value can be calculated exactly from the table
  - Recall t-test ➢We calculate a t-value
  - Based on a distribution, we get the p-value

- How to do Fisher's exact test and calculate its p-value?

	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

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

IV. Cancer genomics overview (Slide: 43-58)

1. What is cancer?

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

2. 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