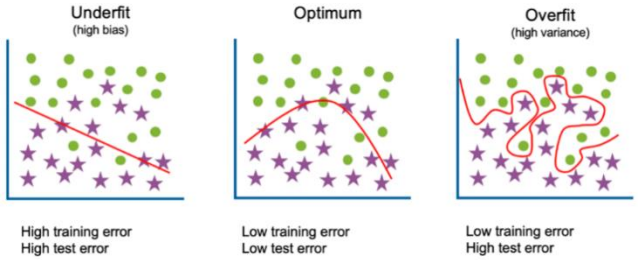
Lec-14 Scribing

Recap

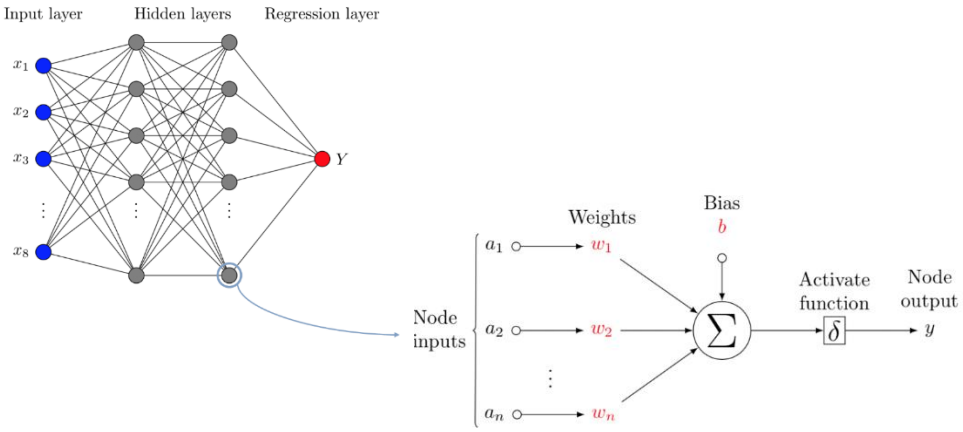
- Logistic Regression (from Lec-10)

New

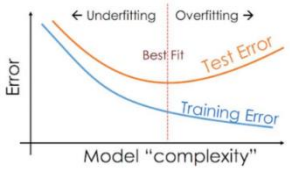
- We can use logistic regression in deep learning for disease screening/classification
  - But the relationship among different variables within the image may be much more complicated than simple linear combination
  - The model capacity is not enough, i.e. Underfitting
  - In practice, we need to overfit the data first



- How to make models with more capacity?
  - To resolve complicated problems
    - Increase the number of nodes
    - Increase the number of layers
    - Add non-linear function
  - Fully-connected layers
    - A general function approximator
    - We can approximate any function (relation) if we have enough nodes and layers
    - Universal approximation theorem



What if the model is much more complex than the problem?

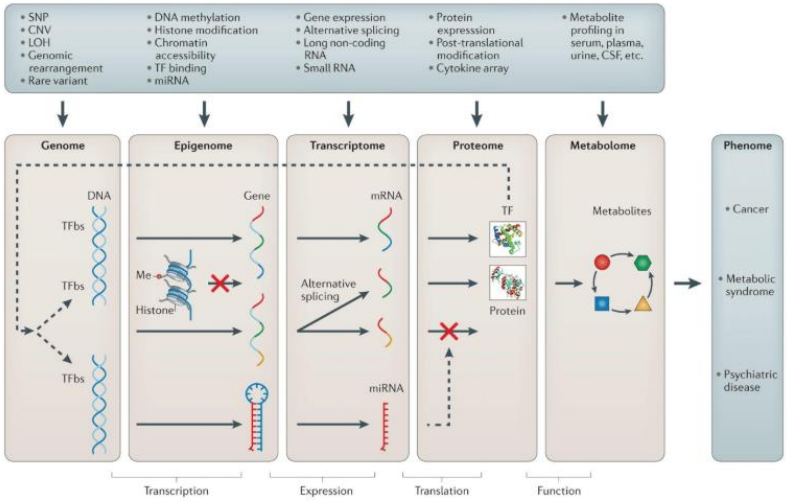


- Typical loss function curves
- Too complex---overfitting
  - The model is too complicated that it may fit the noise in the data
- What is overfitting?
  - Statistically: the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably
  - 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?
  - Train-validation-test split (More preferred in the big data era)
    - Train: 70%
    - Validation: 15%
    - Test: 15%
  - Cross-validation (Can be done if the model is light)
    - 5-fold validation
    - Leave-one-out
    - Reliable evaluation
    - Expensive
- How to deal with overfitting?
  - Data: Too little, not reflect the true distribution
  - Model: Too large, too many useless parameters
  - Connectivity: Too strong, co-adaptation
  - Parameter value range: Too large, model too flexible
  - Training time: Too long, tend to overfitting
  - Model and data; Increase the smoothness

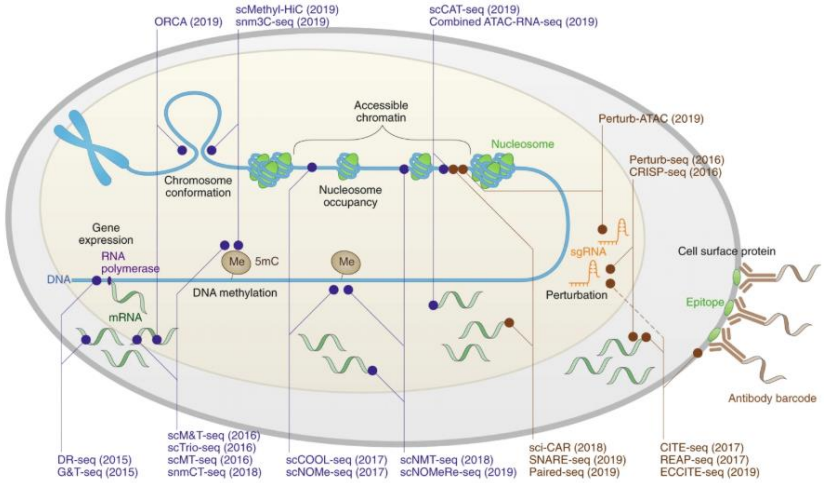
Recap: Why multi-omics (from Lec-1)

- Multi-omics: a growing field (exponential function of time)
- 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

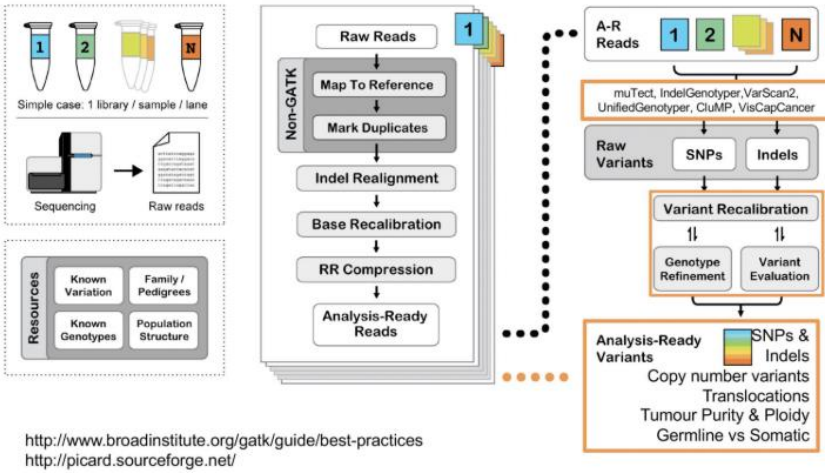
multi-omics



single-cell multi-omics

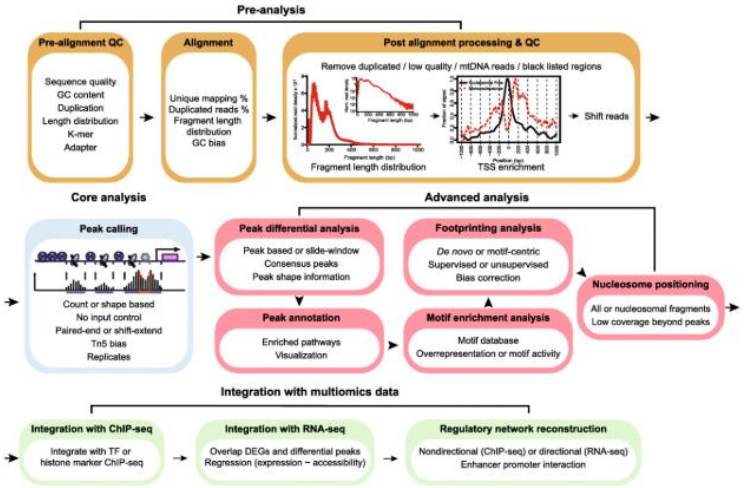


Genome pipeline

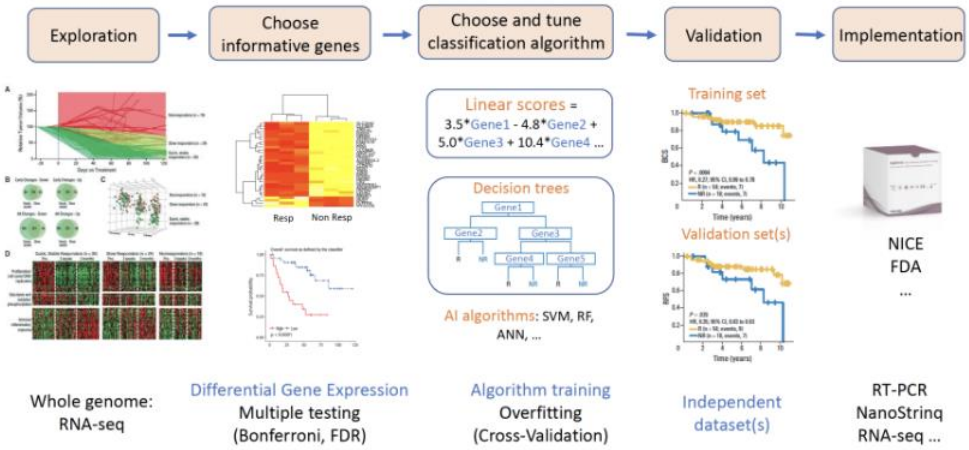


http://www.broadinstitute.org/gatk/guide/best-practices
http://picard.sourceforge.net/

Epigenome pipeline



Transcriptome pipeline



- Multi-omics data analysis can be very tedious nowadays
- But the core techniques are the same
  - Sequence alignment and comparison
  - Dimension reduction and visualization
  - Clustering and classification
- However, one powerful technique is not yet covered in the course
  - To show how confident we are for our claims: statistical testing
- Statistical testing in genomics
  - Variant calling
  - Peak calling
  - Peak differential analysis
  - Differential gene expression analysis
  - Motif enrichment analysis
  - GO enrichment analysis
  - KEGG enrichment analysis
  - Genome-wide association study (GWAS)

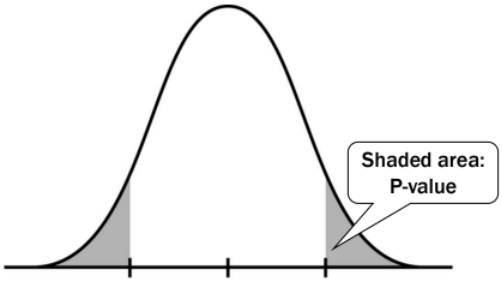
- Differential gene expression analysis
  - Statistical analysis to discover quantitative changes in expression levels between experimental groups
  - For a given gene, whether the gene expression difference is significant, other than due to natural random variation
  - How to compare two sets of values?
- Mean is not enough for comparing the difference
  - Need to consider the variance of two sets of data as well
  - How to compare the gene expressions considering both mean and variance?

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
  - Calculate a test statistic based on the mean and variance of the data
  - Test statistic follows a Student's t-distribution
  - 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?

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

For unpaired, two-tailed t-test



- P-value
  - Usually, the value should be smaller than 0.05
  - We cannot say this gene expressed differently under the two conditions if P-value is large
- Different kinds of t-test
- Paired VS unpaired
  - For paired one, we cannot shuffle 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 kinds of 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)

Gene enrichment analysis

- A biological pathway is a series of interactions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move
  - KEGG pathway database
  - Each pathway contains a set of genes
- By experiments, researchers identified 213 genes associated with type-II diabetes
- Question: how to identify pathways related with type-II diabetes?

Testing association

- A pathway VS type-II diabetes
- If there are related
  - a, d should be large
  - b, c should be small

	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

- How large they should be to say they are related confidently?
  - We need quantitative measure
  - A standard procedure
  - Statistical test for association
  - Fisher's exact test

Fisher's exact test

- 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?

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

Cancer

- What is cancer?
  - a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body
- Why do we want to study cancer?
  - Increasing deaths caused by cancer, which is just lower than those caused by heart disease, i.e. cardiovascular diseases
- How do we study cancer?
  - Cancer is usually believed to be a genomic disease
  - So, we will use genomics/multi-omics methods to study it
  - Genome/Epigenome/Transcriptome/Proteome/Metabolome
- 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
- Neural networks and statistical testing in Python

```
>>> from sklearn.neural_network import MLPClassifier
>>> X = [[0., 0.], [1., 1.]]
>>> y = [0, 1]
>>> clf = MLPClassifier(solver='lbfgs', alpha=1e-5,
...                     hidden_layer_sizes=(5, 2), random_state=1)
...
>>> clf.fit(X, y)
MLPClassifier(alpha=1e-05, hidden_layer_sizes=(5, 2), random_state=1,
              solver='lbfgs')
```

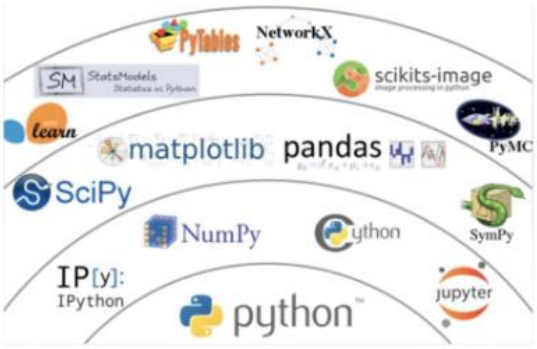
Examples

```
>>> from scipy import stats
>>> rng = np.random.default_rng()
```

Test with sample with identical means:

```
>>> rvs1 = stats.norm.rvs(loc=5, scale=10, size=500, random_state=rng)
>>> rvs2 = stats.norm.rvs(loc=5, scale=10, size=500, random_state=rng)
>>> stats.ttest_ind(rvs1, rvs2)
Ttest_indResult(statistic=-0.4390847099199348, pvalue=0.6606952038870015)
>>> stats.ttest_ind(rvs1, rvs2, equal_var=False)
Ttest_indResult(statistic=-0.4390847099199348, pvalue=0.6606952553131064)
```

- Python's scientific ecosystem



Resources and uncovered topics

- Dive into deep learning
- Dropout
- Generalization
- Out of distribution generalization
- Data distribution
- Multiple testing correction
- How does cancer develop?
- Cancer types