BMEG3105: Data analytics for personalized genomics and precision medicine — Fall 2024 - Lecture 14

Lecturer: Yu LI (李煜) from CSE Liyu95.com, liyu@cse.cuhk.edu.hk

Date: 25 October 2023

Topic: multi-Omics Overview

Lecture Outline

1. Model overfitting

2. Multi-omics overview

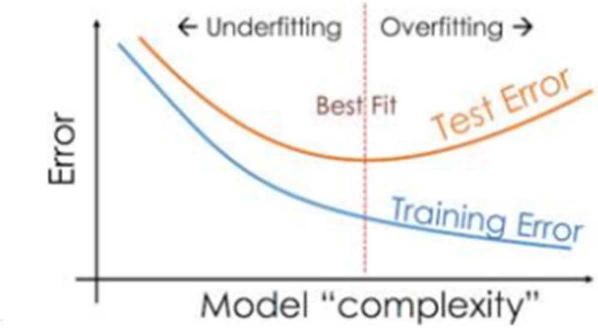
3. Statistical testing

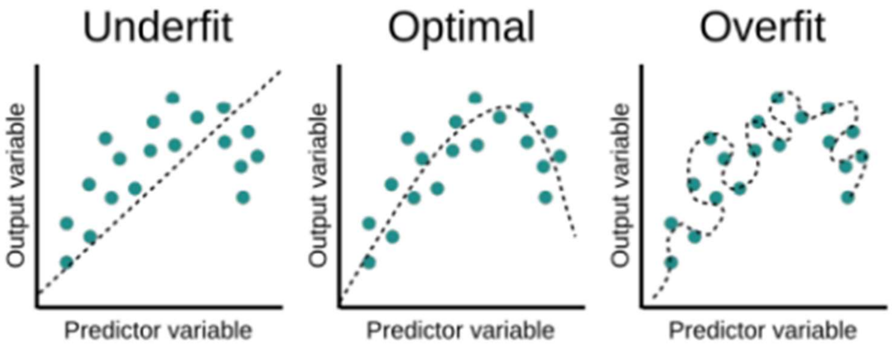
Model Overfitting

Definition

Overfitting occurs when a model learns not only the underlying patterns in the training data but also the noise, leading it to perform well on the training set but poorly on unseen data.

- **Statistical Perspective**: It is the production of an analysis that corresponds too closely to a particular dataset, failing to generalize to new data.
- **Machine Learning Perspective**: The model is more complex than necessary for the problem at hand, allowing it to excel on the training dataset while underperforming on the test dataset.



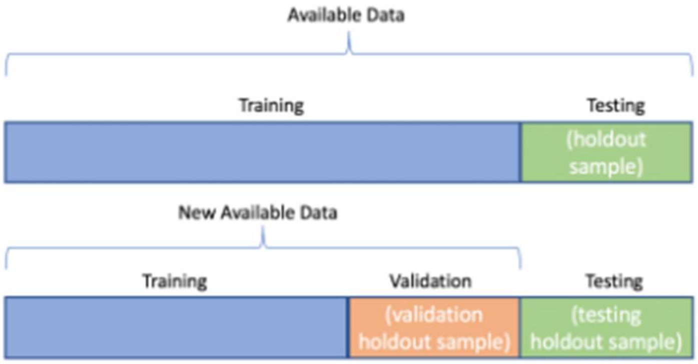


Evaluation of Overfitting

To evaluate and detect overfitting, the following methods are commonly used:

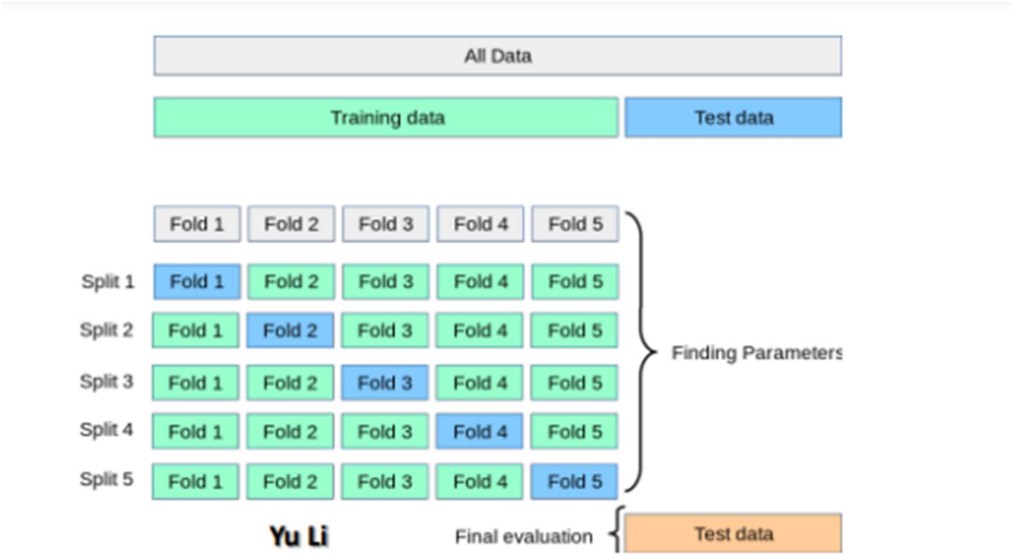
1. Train-Validation-Test Split

- **Split Ratio**: 70% training, 15% validation, 15% testing.
- **Purpose**: The test set must remain untouched during training to accurately assess model performance. A significant gap between training and testing accuracy indicates potential overfitting.



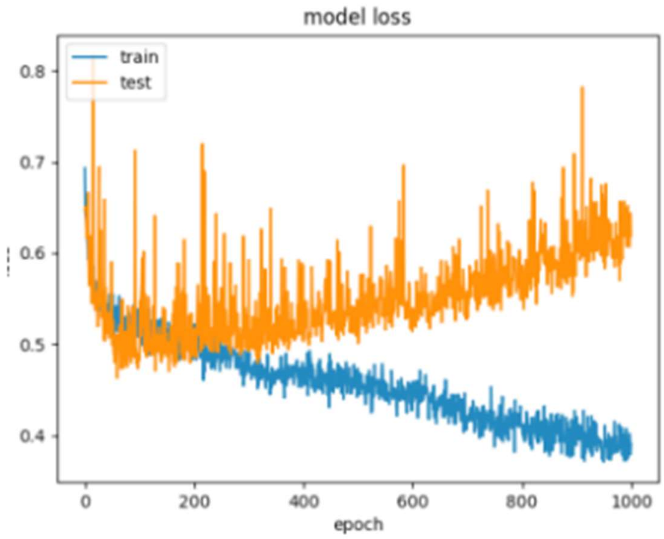
2. Cross-Validation

- **Techniques**:
  - 5-fold validation
  - Leave-one-out validation
- **Pros and Cons**: Provides a more reliable evaluation but is computationally expensive.



3. Loss Function Analysis

- **Training vs. Validation Loss**: A graph showing training loss decreasing while validation loss increases over epochs signals overfitting.
- **Early Stopping**: Monitor losses to halt training before overfitting occurs.



4. Performance Metrics

- Evaluate using precision, recall, and F1 score. Ensure all metrics are reasonable and highlight potential issues with model bias towards the majority class.

Sources of Over-Complexity

Overfitting can stem from several factors:

1. **Insufficient Data**: Small datasets may not accurately reflect true distributions, leading to misleading model learning.
2. **Complex Model Architecture**: A model with excessive parameters may capture irrelevant noise.
3. **Strong Connectivity**: Excessive reliance on specific features can cause co-adaptation among model nodes.
4. **Large Parameter Value Range**: Excessive flexibility in a model can lead to overfitting.
5. **Extended Training Duration**: Long training times can cause models to memorize noise rather than learn patterns.

Techniques to Mitigate Overfitting

Several strategies can be employed to address overfitting:

1. **Increase Dataset Size**: Augment the training dataset to better reflect the true distribution.
2. **Data Augmentation**: Introduce controlled noise to inputs or outputs to enhance dataset diversity without compromising quality.
3. **Regularization**: Implement techniques like L1/L2 regularization to penalize large weights, reducing model complexity.
4. **Early Stopping**: Stop training before the model starts memorizing noise.
5. **Simplifying the Model**: Reduce complexity by pruning decision trees, minimizing neural network parameters, or applying dropout during training.

Multi-Omics

Definition

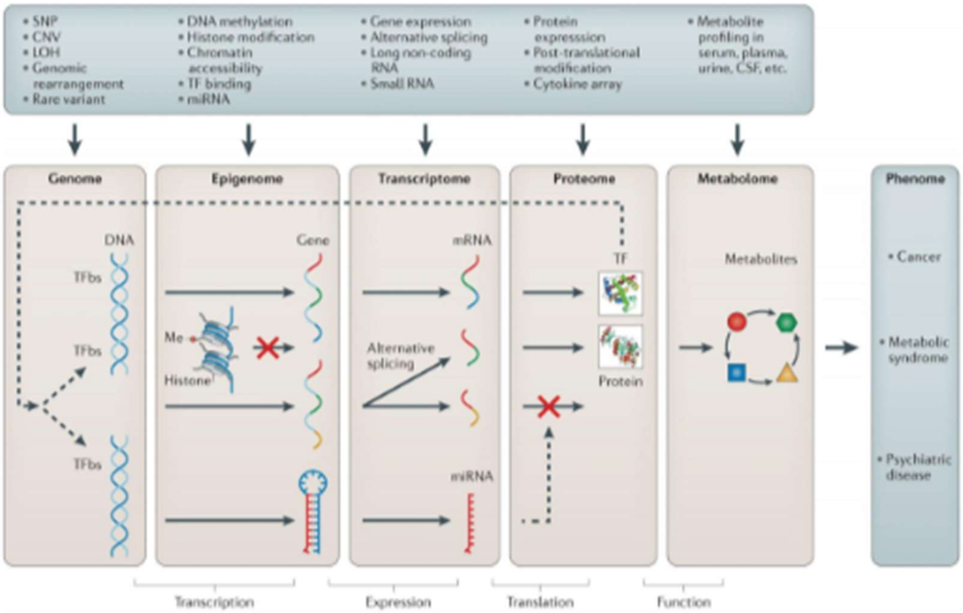
Multi-Omics refers to a comprehensive, longitudinal big data approach aimed at precision health by analyzing multiple biological datasets collectively. It involves the characterization and quantification of various pools of biological molecules that reflect the structure, function, and dynamics of organisms.

Omics Overview

What is Omics?

Omics encompasses various biological fields focused on the collective analysis of biological molecules. These analyses help understand the interactions and dynamics within an organism, emphasizing the need to consider all omics together due to their interconnectedness.

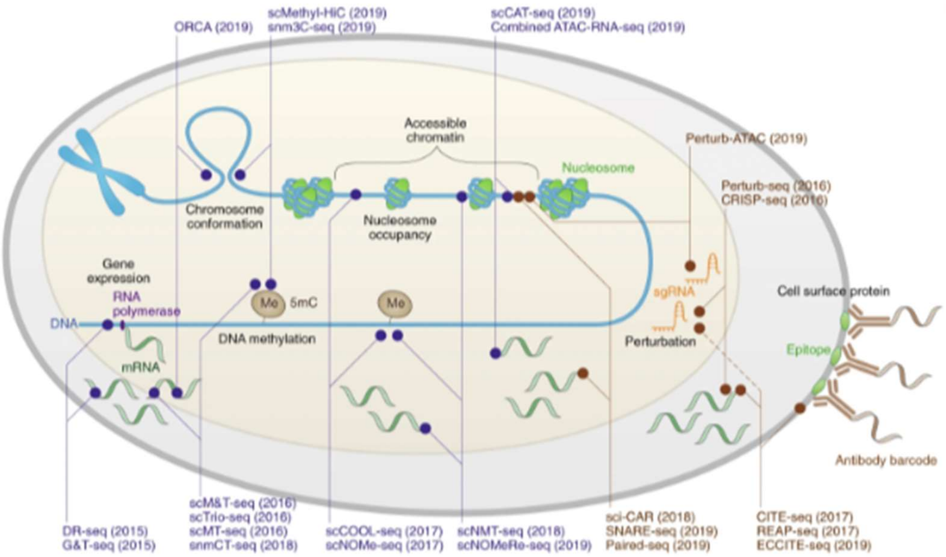
- **Transcriptome**: Involves transcribing DNA into RNA, including processes such as alternative splicing, which modifies immature mRNA into its mature form.
- **Proteome**: Refers to the total set of proteins expressed by a genome, playing critical roles in regulating biological functions.



Nature Reviews | Genetics

Single-Cell Multi-Omics

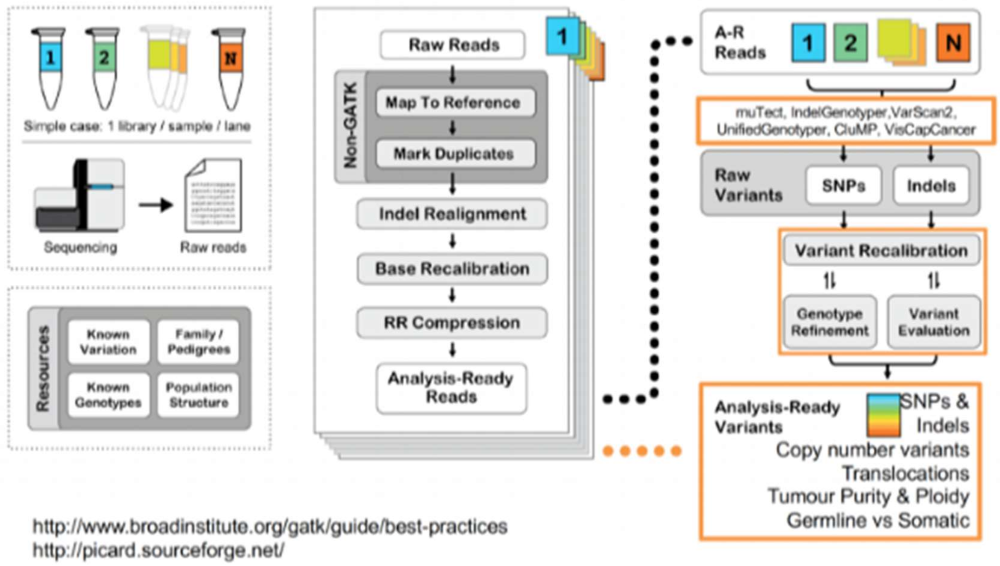
This approach enables the analysis of multiple cell types at a granular level, facilitating exhaustive investigations into cellular heterogeneity and the interactions between different cell types.



Types of Omics

1. **Genome Pipeline**:

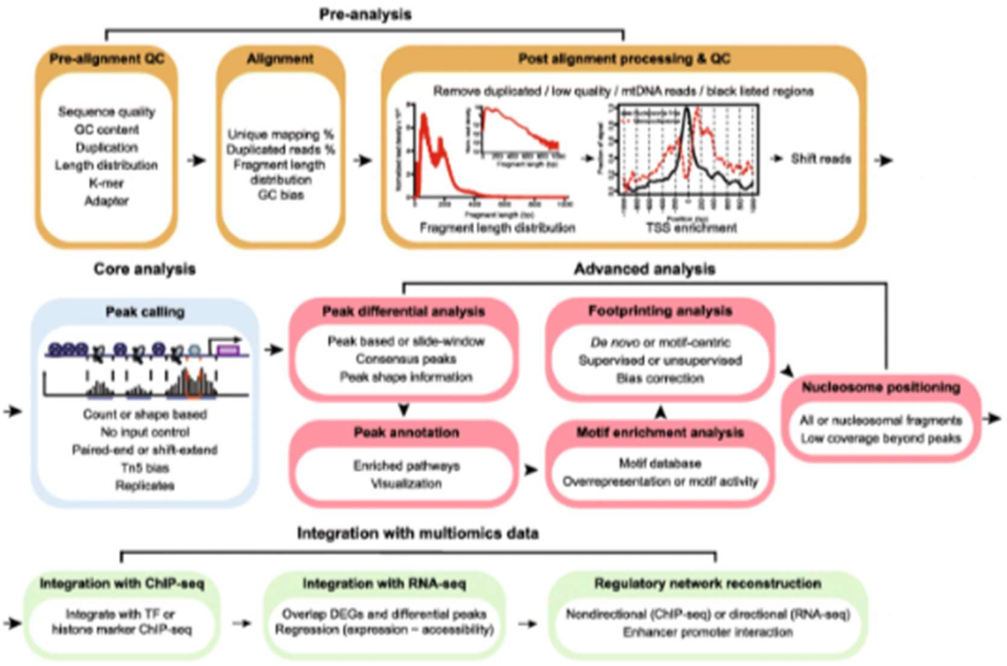
   - Analyzes genetic variations and their associations with phenotypic differences among individuals.



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

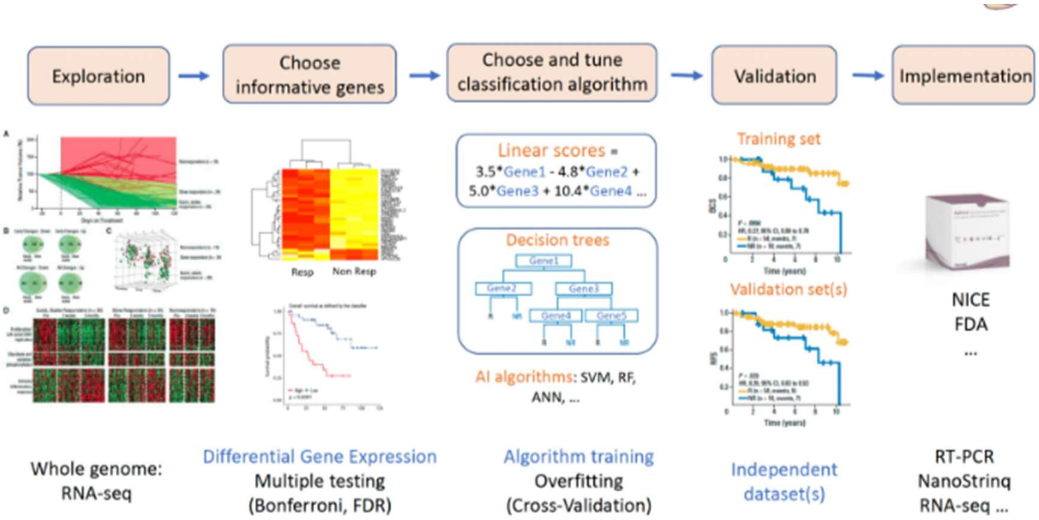
2. **Epigenome Pipeline**:

   - Investigates reversible modifications that affect gene expression without altering DNA sequences, leading to different phenotypes despite similar genetic backgrounds.


Pre-analysis
Pre-alignment QC
Sequence quality
GC content
Duplication
Length distribution
K-mer
Adaptor
Alignment
Unique mapping %
Duplicated reads %
Fragment length distribution
GC bias
Post alignment processing & QC
Remove duplicated / low quality / mtDNA reads / black listed regions
Fragment length distribution
TSS enrichment
Shift reads
Core analysis
Advanced analysis
Peak calling
Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates
Peak differential analysis
Peak based or slide-window
Consensus peaks
Peak shape information
Footprinting analysis
De novo or motif-centric
Supervised or unsupervised
Bias correction
Nucleosome positioning
All or nucleosomal fragments
Low coverage beyond peaks
Peak annotation
Enriched pathways
Visualization
Motif enrichment analysis
Motif database
Overrepresentation or motif activity
Integration with multiomics data
Integration with ChIP-seq
Integrate with TF or histone marker ChIP-seq
Integration with RNA-seq
Overlap DEGs and differential peaks
Regression (expression ~ accessibility)
Regulatory network reconstruction
Nondirectional (ChIP-seq) or directional (RNA-seq)
Enhancer promoter interaction

3. **Transcriptome Pipeline**:

   - Focuses on understanding the expression levels of genes and how they are regulated.



4. **Proteome Pipeline**:
   - Studies the entire set of proteins in a cell, tissue, or organism, examining their functions and interactions.
5. **Metabolome and Other Omics**:
   - Expands to include analyses of metabolites, lipidome, phosphoproteome, and glycoproteome, among others.

Key Messages

- Multi-omics data analysis can be complex and time-consuming, but the core analytical techniques remain consistent.
- Fundamental techniques include:
  - **Sequence Alignment and Comparison**: To identify similarities and differences across datasets.
  - **Dimension Reduction and Visualization**: For simplifying complex data into interpretable formats.
  - **Clustering and Classification**: To group similar biological samples or conditions based on their omic data.

Statistical Testing

Overview

Statistical testing is crucial for discovering quantitative changes in expression levels between experimental groups. It assesses whether observed differences in gene expression are significant or merely due to natural variation.

1. Differential Gene Expression Analysis

- Aims to identify significant changes in gene expression levels between different experimental groups.
- Determines if the observed differences are statistically significant, beyond random variation.

2. T-Test

Definition

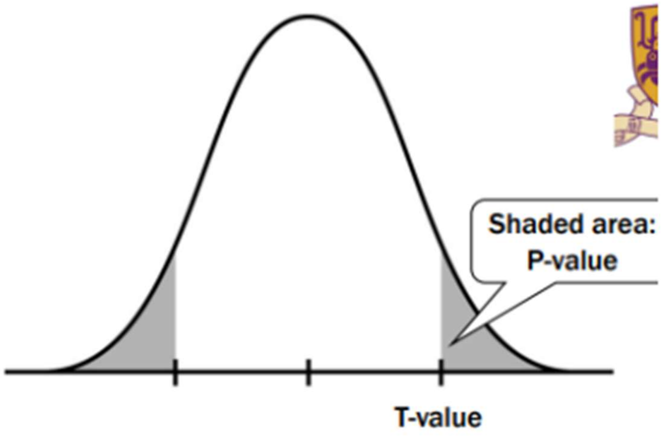
A T-test is a standard statistical procedure used to determine if there is a significant difference between the means of two groups.

Purpose

- To assess whether the difference in gene expression between two conditions is significant.

Procedure

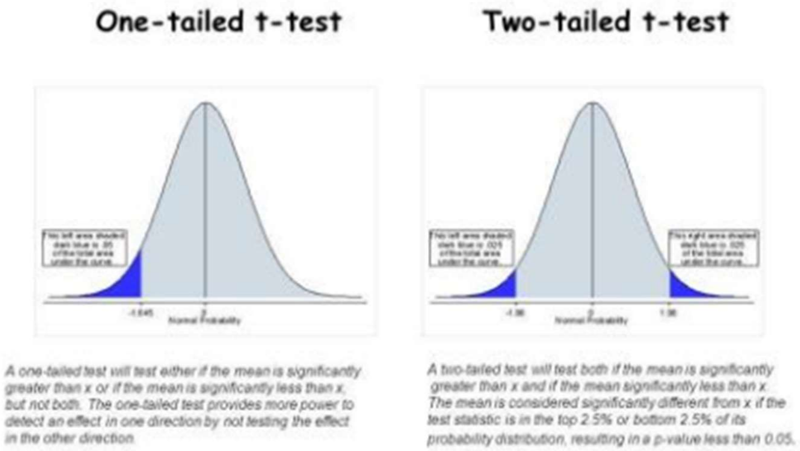
1. **Calculate the Test Statistic**:
   - Based on the mean and variance of the data.
   - The test statistic follows a Student's t-distribution.
2. **Calculate the P-Value**:
   - Represents the probability that the observed result occurred by chance.
   - A smaller p-value indicates greater confidence in the significance of the result.
   - **Example**: For an unpaired two-tailed t-test, a p-value smaller than 0.05 suggests that the two gene expressions are significantly different.



$$T=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}=\frac{15.09-13.00}{\sqrt{\frac{146.69}{11}+\frac{18.22}{10}}}=\frac{2.09}{3.894}=0.54$$

Types of T-Tests

- **One-Tailed Test**: Tests for a specific direction of difference (greater, larger, smaller).
- **Two-Tailed Test**: Tests for any difference (different or the same).



A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both. The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x. The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

Conclusion

- If the p-value is less than 0.05, we can conclude that the gene expression differs significantly under the two conditions.

3. Gene Enrichment Analysis

- **Objective**: Identify pathways associated with specific conditions (e.g., Type II diabetes).

- Involves assessing the relationship between genes and biological pathways.
- A contingency table is created to analyze the relationship:
    - High counts in cells 'a' and 'd' indicate a potential relationship.
    - Low counts in cells 'b' and 'c' further support this relationship.

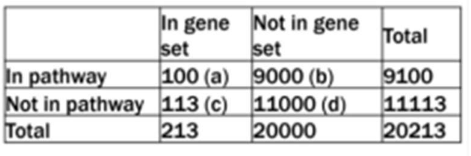
4. Testing Association

- Examines whether a given pathway is related to specific genes (e.g., those associated with Type II diabetes).
- The goal is to determine if there is a significant association based on the distribution of related and unrelated genes.

Fisher's Exact Test

- A statistical test used to analyze contingency tables.
- Calculates the exact p-value based on the counts in the table.
- If the p-value (e.g., p = 0.5802) is greater than 0.05, the pathway is not related to Type II diabetes.

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



	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

5. Summary of Statistical Testing

5.1 Statistical Analysis

- Aims to discover quantitative changes in gene expression levels between experimental groups.
- Mean and variance alone may not adequately represent the differences across groups.

5.2 T-Test

- Checks for significant differences between data sets.
- Steps:
    1. Calculate a test statistic following the Student's t-distribution.
    2. Compare the test statistic with the p-value.

5.3 Gene Enrichment Analysis

- Identifies biological pathways through contingency tables.
- High counts in specific cells indicate a relationship, while low counts in others suggest independence.