



# Introduction to Statistics in Genomics

Jenny Wu, Ph.D.

Director of Bioinformatics

Genomics Research and Technology Hub  
Chao Family Comprehensive Cancer Center

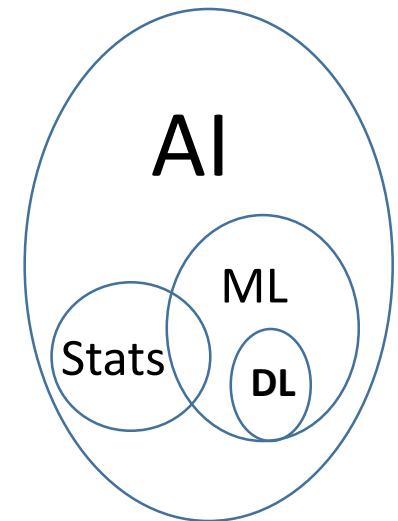
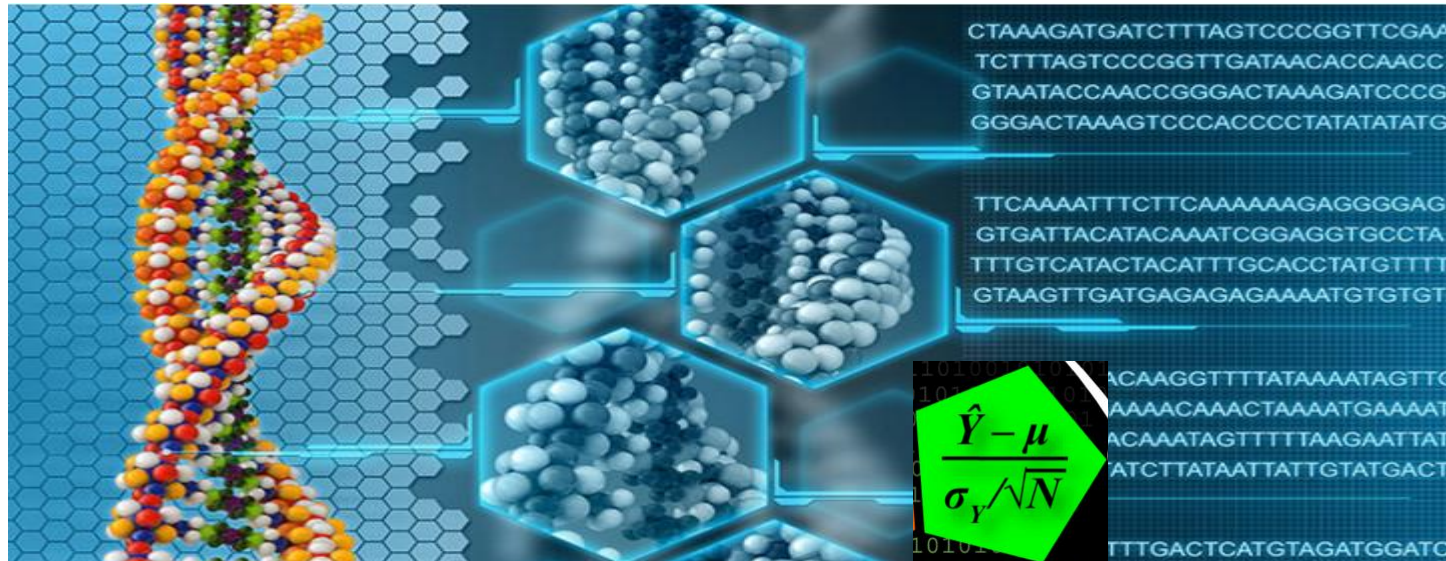
UC Irvine

# Outline

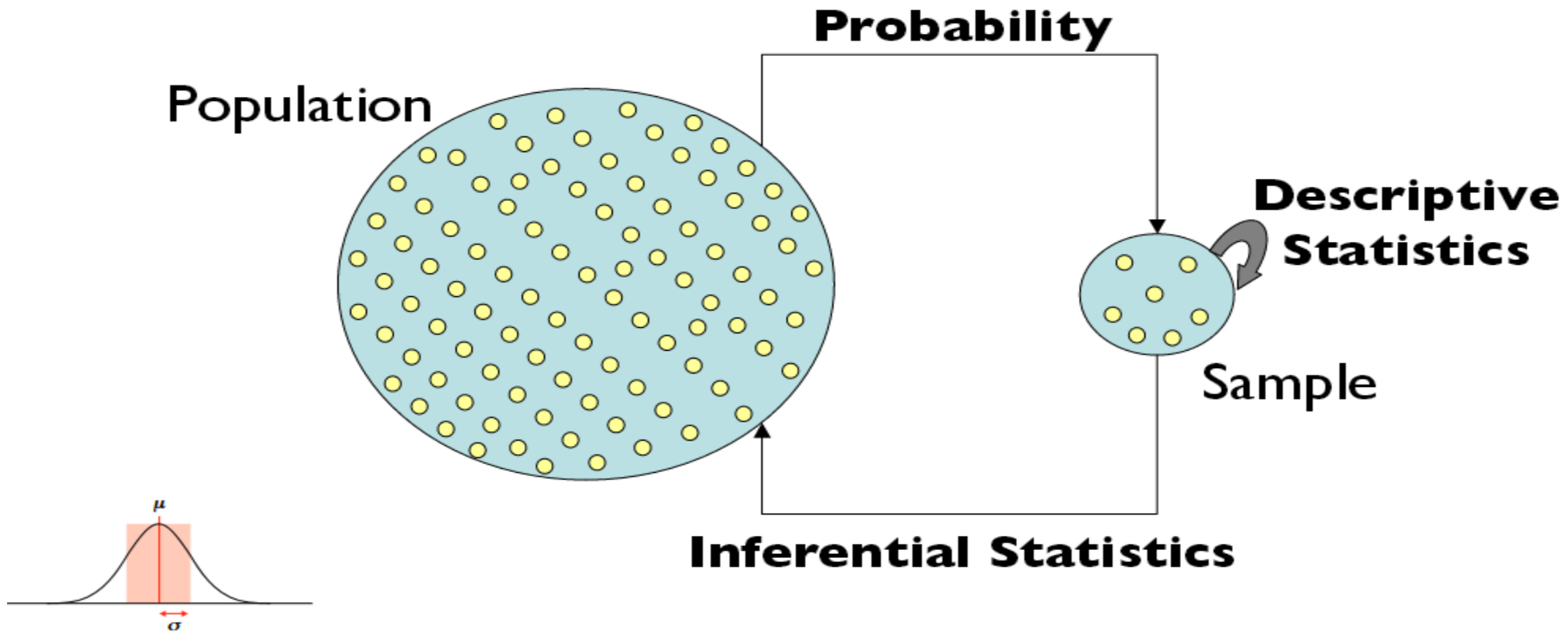
- Goals : Practical guide to statistical analysis of genomic data
- Statistical concepts
  - Central Dogma of Statistics
  - Basic concepts
    - Random variable and probability distribution
    - Hypothesis Testing
    - P value
    - Linear models
- Statistics applications in omics data
  - Generalized linear models (GLMs) in count data
  - Multiple hypothesis testing for omics
  - *Clustered data and batch effect*
  - *Multi Omics integration*
- Pathway Analysis
- Statistics vs Machine learning (ML)
- *Hands on session*

# Why statistics in Omics

- **Omics** = Massive amount of Data
- **Statistics** is fundamental in genomics because it is integral in the **design, analysis** and **interpretation** of experiments.

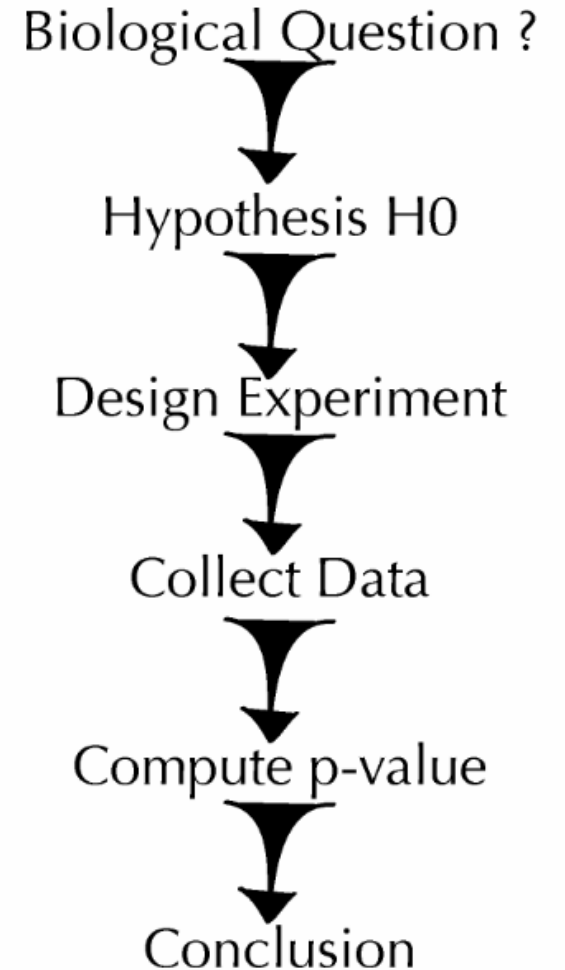


# “Central Dogma” of Statistics



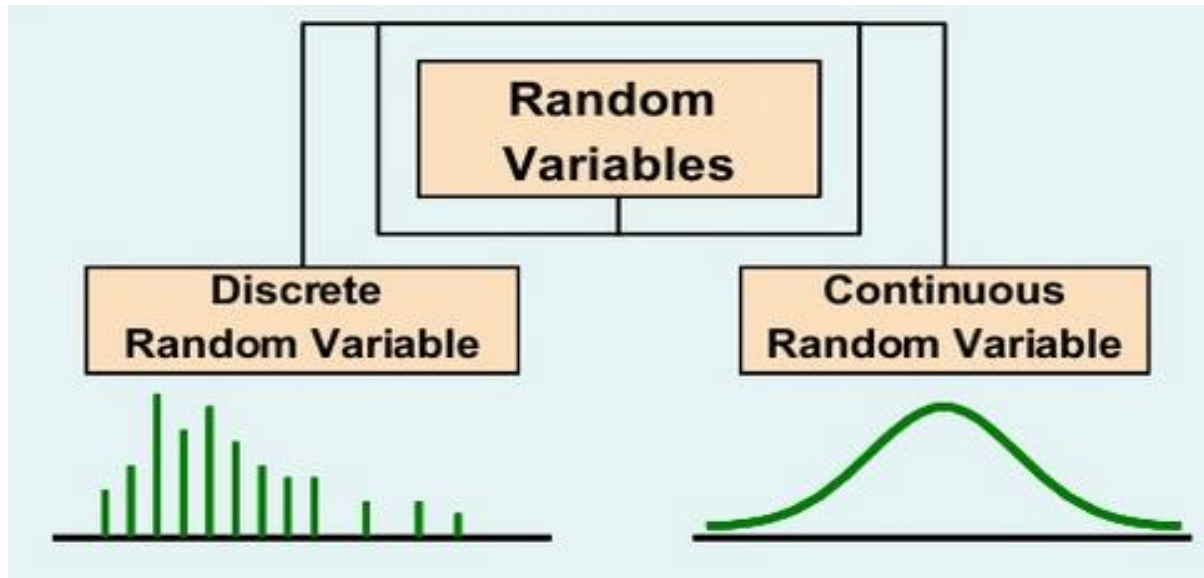
# Basic Concepts

- Units: the basic objects on which the experiment is done. Sampling, experimental vs observational
- Variable: a measured characteristic of a unit
- Treatment: any specific experimental condition applied to the units
- Hypothesis: A hypothesis is a statement about a parameter of interest. Hypothesis testing is formalized to make a decision between rejecting or not rejecting a null hypothesis on the basis of a set of experimental observations and measurements.



# Random Variables and Prob Distribution

- A **Discrete** r.v. has a countable number of possible outcomes. *e.g. genotype of a SNP, read counts in RNA-seq etc.* Categorical or Ordinal
- A **Continuous** r.v. can adopt any value in an interval of numbers. *e.g. height, weight, microarray measurements of gene expression level etc.*



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} .$$

$$P(X) = \frac{e^{-\lambda} \lambda^x}{x!}$$

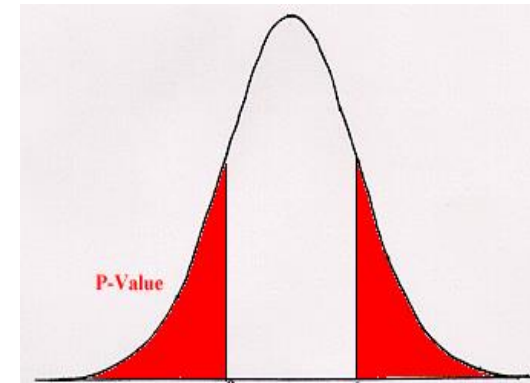
# Hypothesis Testing

- The intent of hypothesis testing is formally examining two opposing conjectures  $H_0$  (null hypothesis) and  $H_A$  (alternative hypothesis)
- Steps:
  - Set up  $H_0$  and  $H_A$  to state the assumption to be tested.  $H_A$  is the opposite of null. Is generally to be believed by the researcher.
  - Select a test statistic ( $t$ ,  $z$  etc)
  - Set up decision rule (e.g.  $\alpha = 0.05$ )
  - Compute test statistic
  - Draw conclusion and summarize significance



# P value

- Calculate the test statistic from the sample data
- Convert test statistic to a p value by comparing its value to the null distribution: distribution of test statistic under  $H_0$ .
- P value is the probability of observing as or more extreme value by chance based on the null distribution.

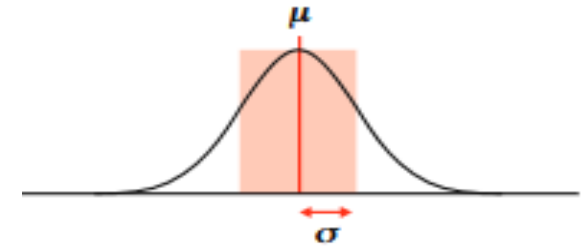


$$p \text{ value} \leq \alpha \rightarrow \text{Reject } H_0$$
$$p \text{ value} > \alpha \rightarrow \text{Do not reject } H_0$$



# Linear Models (LM) and Hypothesis Testing

- The most widely used models in statistics
- $X$ : predictors, explanatory variable
- $Y$ : response variable, dependent variable
- Design (model) matrix  $X$ , contrast



$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2)$$

$$y = X\beta + e,$$

# Generalized Linear Models (GLM) in Count Data

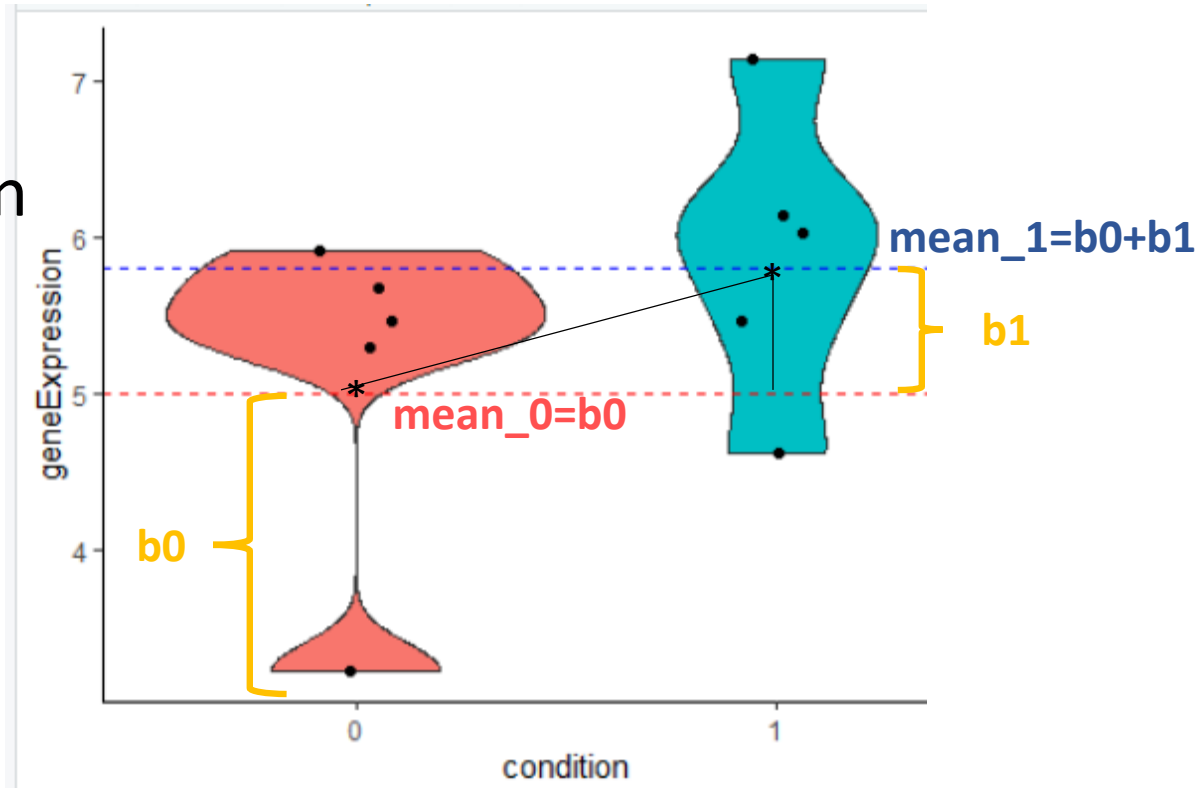
- Response variable is assumed to follow an exponential family distribution with mean
- 3 components: Random, Systematic, and Link Function
  - Random component: Identifies dependent variable ( $Y$ ) and its probability distribution
  - Systematic Component: Identifies the set of explanatory variables ( $X_1, \dots, X_k$ )
  - Link Function: Identifies a function of the mean that is a linear function of the explanatory variables

$$g(\mu) = \alpha + \beta_1 X_1 + \dots + \beta_k X_k$$

# Classic Framework for Omics Data Analysis

## Generative Approach: Regression

Y: gene expression  
X: condition



**Condition:**  $x = 0$  (*control*) or  $1$  (*diseased*)

**Gene Expression:**  $y = \log_2(q_i) = b_0 + b_1x$   
 $y = \log_2(q_i) = b_0 + b_1x_1 + b_2x_2$

# GLMs in RNA-seq: DESeq2 Implementation

$$K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$



$$\mu_{ij} = s_j q_{ij}$$

$$\log_2(q_{ij}) = x_{j*} \vec{\beta}_i$$

$K_{ij}$	counts of reads for gene $i$ , sample $j$
$\mu_{ij}$	fitted mean
$\alpha_i$	gene-specific dispersion
$s_j$	sample-specific size factor
$q_{ij}$	parameter proportional to the expected true concentration of fragments
$x_{j*}$	the $j$ -th row of the design matrix $X$
$\vec{\beta}_i$	the log fold changes for gene $i$ for each column of $X$

# Design Matrix $X$

$$\begin{array}{cc} X & * & \beta = \log_2(\mu) \\ \begin{array}{c} \boxed{1 \ 0} \\ 1 \ 0 \\ 1 \ 0 \\ 1 \ 0 \\ 1 \ 1 \\ 1 \ 1 \\ 1 \ 1 \\ 1 \ 1 \end{array} & * & \begin{array}{c} \boxed{b_0} \\ b_1 \end{array} = \begin{array}{c} b_0 \\ b_0 \\ b_0 \\ b_0 \\ b_0 + b_1 \\ b_0 + b_1 \\ b_0 + b_1 \\ b_0 + b_1 \end{array} \end{array}$$

 intercept       slope

# Hypothesis Testing in RNA-Seq

Null hypothesis ( $H_0$ )

- The experimental condition  $r$  has no influence on the expression of the gene under consideration:

$$\mu_{\rho_1} = \mu_{\rho_2}$$

Alternative hypothesis ( $H_0$ )

$$\mu_{\rho_1} \neq \mu_{\rho_2}$$

# Hypothesis Testing with GLM

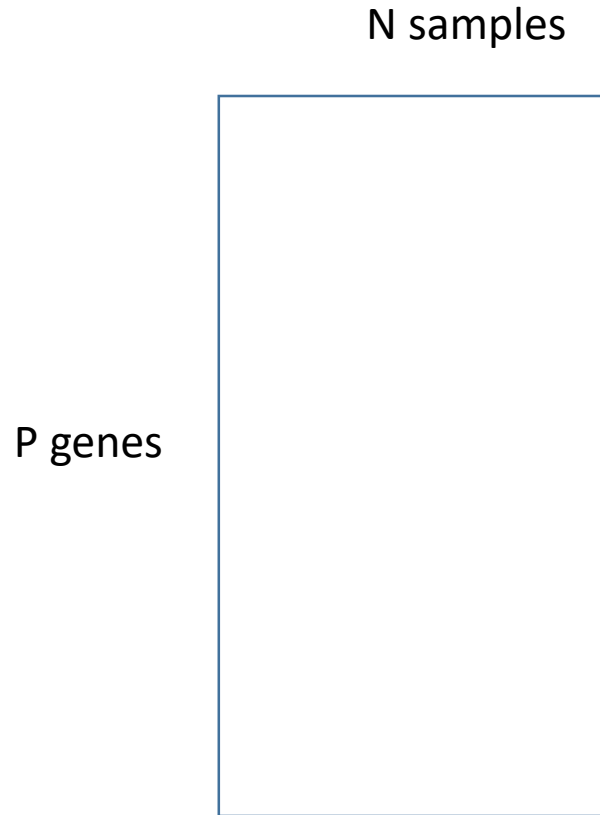
- $H_0: \beta_1 = 0$
- Likelihood Ratio Test

$$D = -2 * \log \left( \frac{L_0}{L_{alt}} \right)$$

Under  $H_0$ ,  $D \sim \chi^2$  and p value can be calculated using  $\chi^2$  distribution.



# Why adjusted p value in Genomics



- Lots of data in genomics that have lots of hypothesis tests.
- In RNA-seq we are doing  $p$  simultaneous tests!  $H_1, H_2, H_3, \dots, H_p$
- For a **10k** gene experiment, a standard p value cutoff 0.05 will give **500** DEGs by chance.

# Multiple Hypothesis Testing

- Simultaneous testing for thousands of genes
- p-values not sufficient to control false positive rate
- Control Family Wise Error Rate (FWER): Bonferroni's solution is too conservative for very high-dimensional data.
- Control False Discovery Rate (FDR) with Benjamini-Hochberg method.

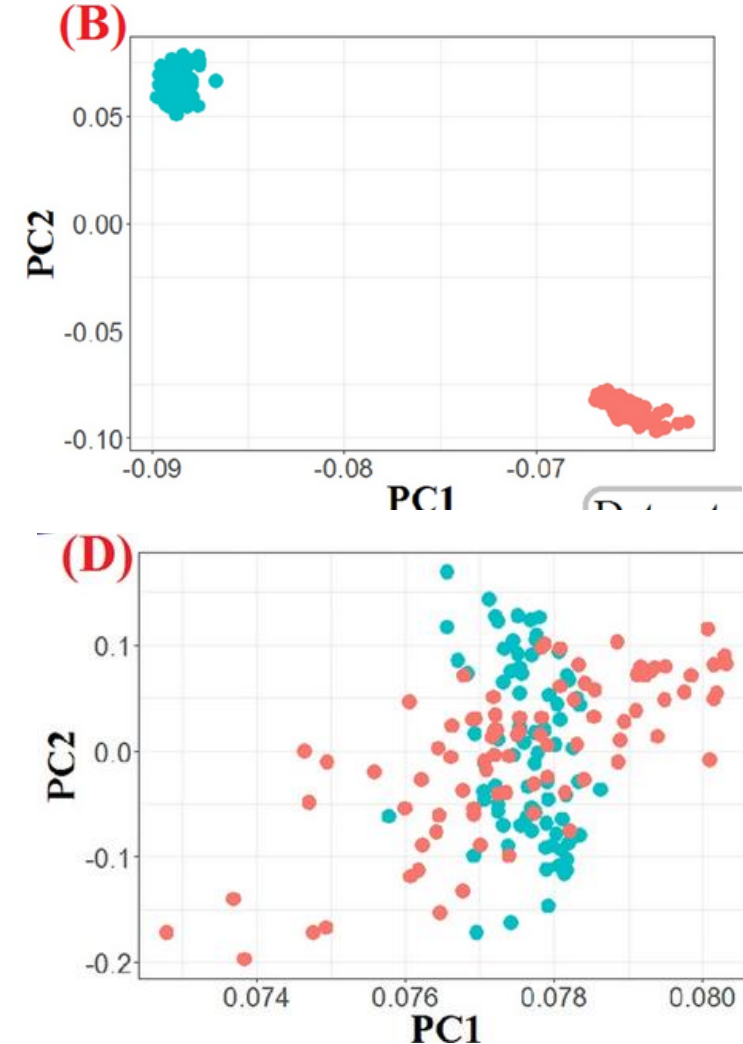
# Clustered Data and Mixed Effect Models

- Omics data are sometimes clustered or collected with repeated measure
- It is important to take data dependence into account
- Implemented in R package *nlme* and *lme4*

$$Y_{ij} = \beta_0 + x_{ij,1}\beta_1 + \dots + x_{ij,4}\beta_4 + u_i + \varepsilon_{ij},$$

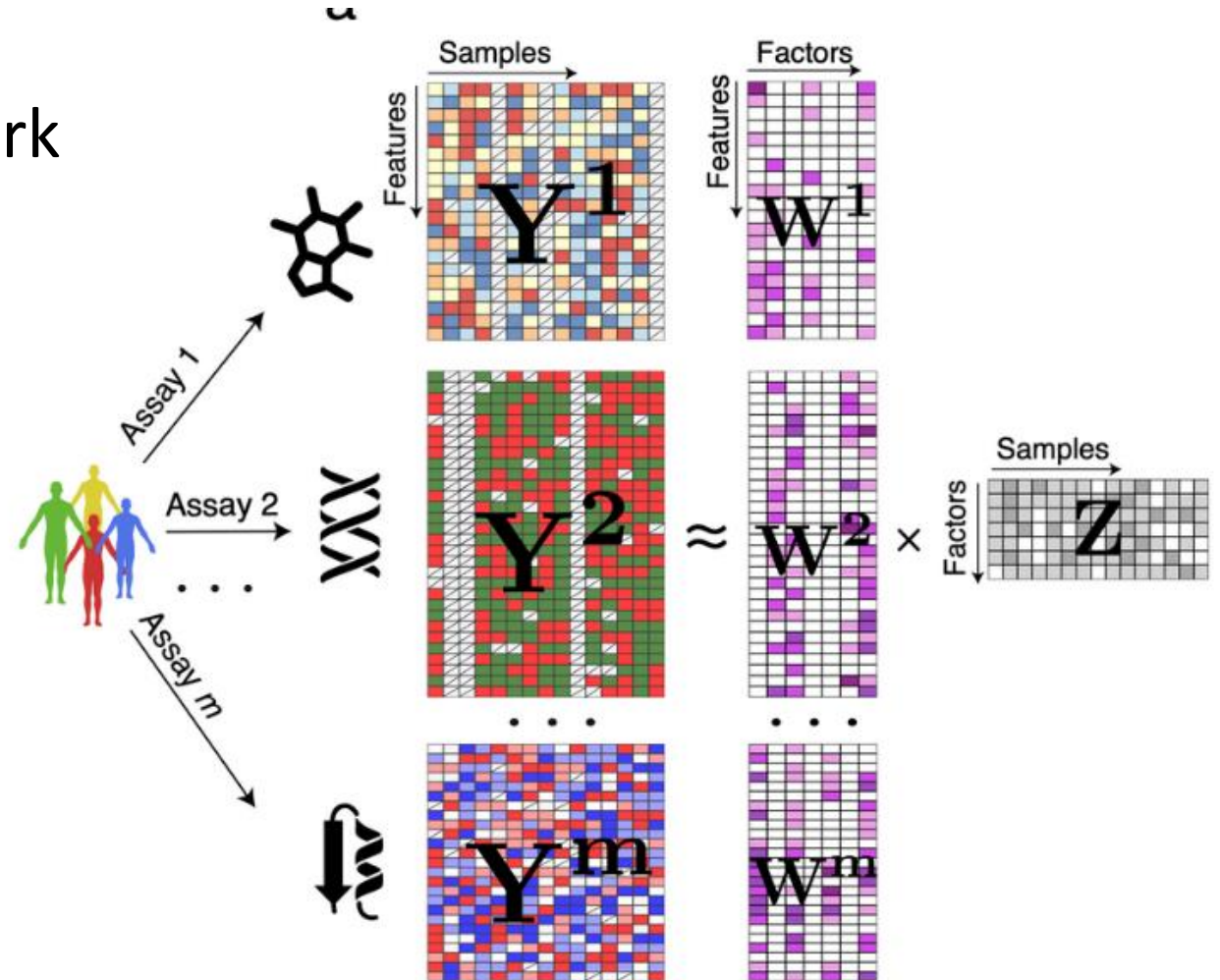
# Batch Effect in Omics Data

- Systematic variations or biases introduced into the data due to technical factors during sample processing or analysis
- When batch is known, use covariates (*limma*, *DESeq*, *Combat*). When unknown, use surrogate variable analysis (*sva* or *RUVseq*)



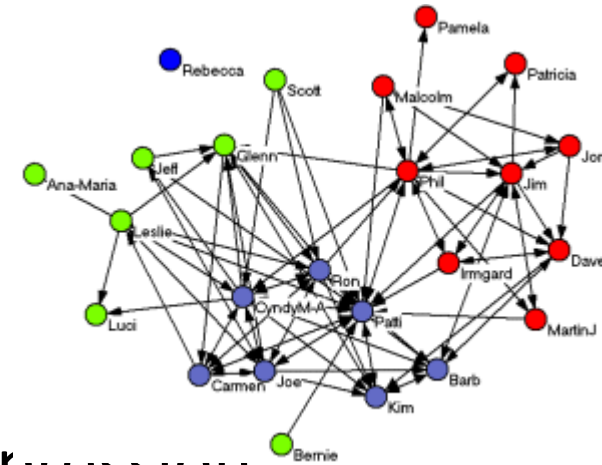
# Statistics in Multi Omics Integration

- MOFA: a factor analysis model that provides a general framework for the integration of multi-omic data sets in an unsupervised fashion
- Other methods such as LIGER, Seurat, and deep learning based tools are popular.



# Why Pathway Analysis

- Logical next step in any high throughput experiment
- Goal: to characterize biological meaning of joint changes in gene expression
- Why? Often sets of genes doing related functions are changed



# Pathway and Network Analysis

## **Pathway Analysis Methods:**

- **Functional category over representation:** discrete test for significance (*BiNGO, David, Gorilla, IPA etc*)
- Continuous test (*GSEA, PAGE*)
- Signaling Pathway Impact Analysis (*iPathway Guide*)

## **Network Analysis:** (*WGCNA, Cytoscape etc*)



# Functional Category Enrichment

- Discrete tests: enrichment for groups in gene lists

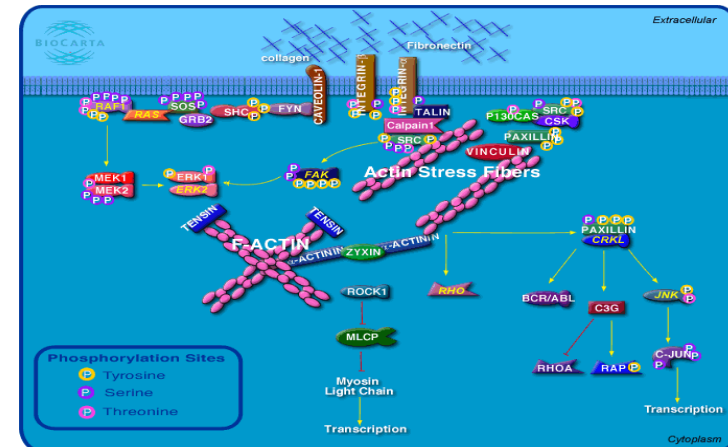
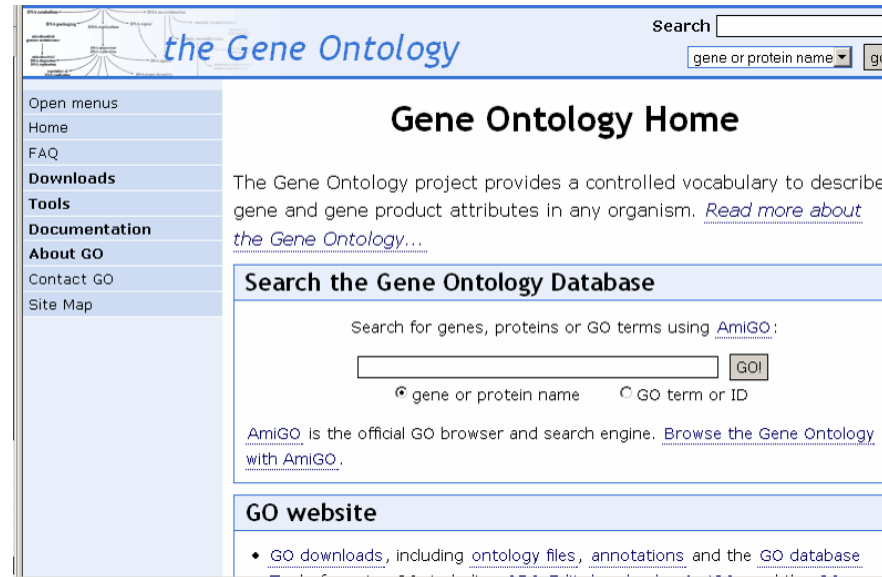
- Select gene list at some predefined cutoff
- For each gene list and functional category cross-tabulate to get a 2X2 contingency table
- Test for significance using Fisher's exact test
- FDR correction for multiple hypothesis testing

	Differentially expressed	Not differentially expressed	total
In the pathway	a	b	a+b
Not in the pathway	c	d	c+d
total	a+c	b+d	n

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

# Functional Categories in Pathway Analysis

- Gene Ontology
  - Biological Process
  - Molecular Function
  - Cellular Location
- Pathway Databases
  - KEGG
  - BioCarta
  - Broad Institute
  - *Commercial knowledge bases such as IPA*
- Other
  - Transcription factor targets
  - Protein complexes
  - *Self-Defined*

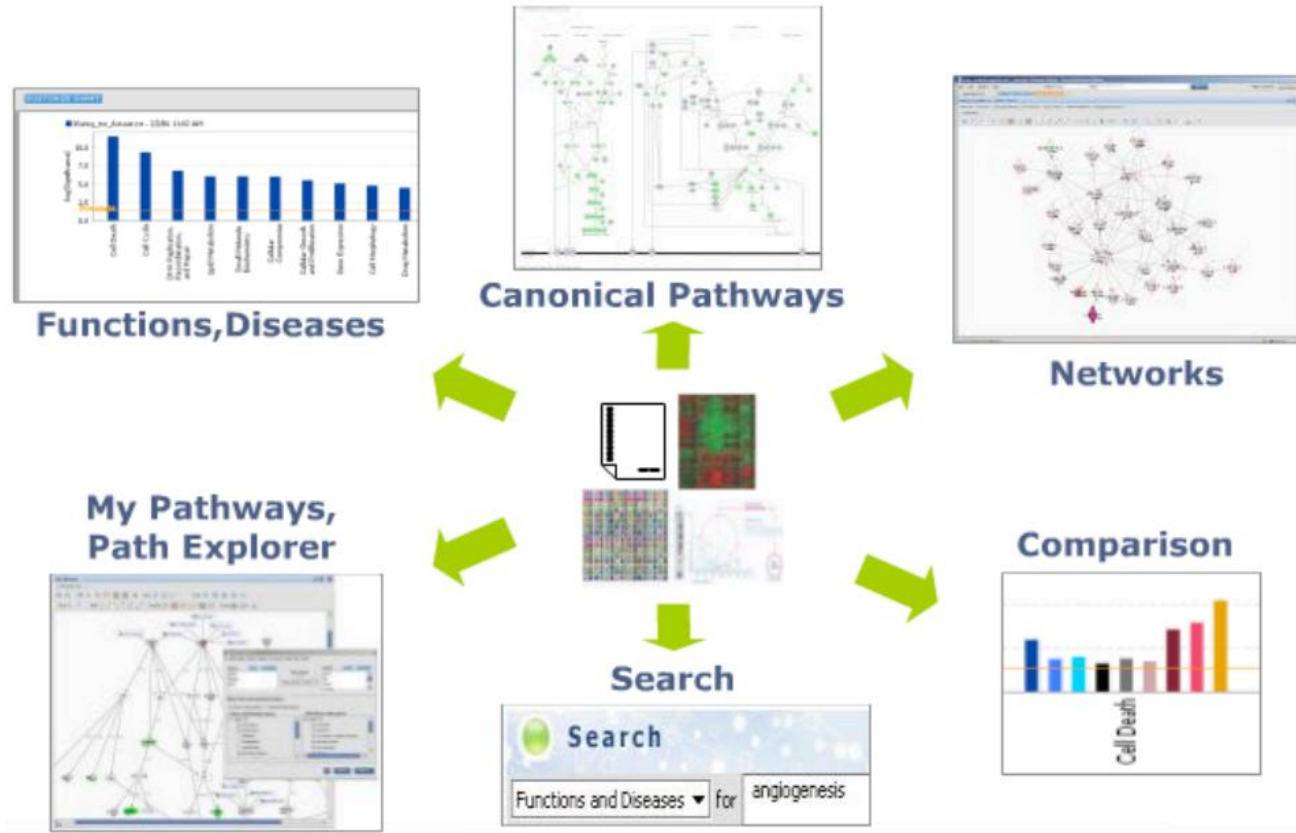


# Commercial and Open Source Pathway Analysis Software

- **GeneGo/MetaCore** ([www.genego.com](http://www.genego.com))
- **Ingenuity Pathway Analysis** ([www.ingenuity.com](http://www.ingenuity.com))
- **Pathway Studio** ([www.ariadnegenomics.com](http://www.ariadnegenomics.com))
- **GenMAPP** ([www.genmapp.com](http://www.genmapp.com))
- **WikiPathways** ([www.wikipathways.org](http://www.wikipathways.org))
- **cPath** ([cbio.mskcc.org/cpath](http://cbio.mskcc.org/cpath))
- **BioCyc** ([www.biocyc.org](http://www.biocyc.org))
- **Pubgene** ([www.pubgene.org](http://www.pubgene.org))
- **PANTHER** ([www.pantherdb.org](http://www.pantherdb.org))
- **WebGestalt** ([bioinfo.vanderbilt.edu/webgestalt/](http://bioinfo.vanderbilt.edu/webgestalt/))
- **ToppGene Suite** ([/toppgene.cchmc.org/](http://toppgene.cchmc.org/))
- **DAVID** ([david.abcc.ncifcrf.gov/](http://david.abcc.ncifcrf.gov/))
- **Pathway Painter** ([pathway.painter.gsa-online.de/](http://pathway.painter.gsa-online.de/))

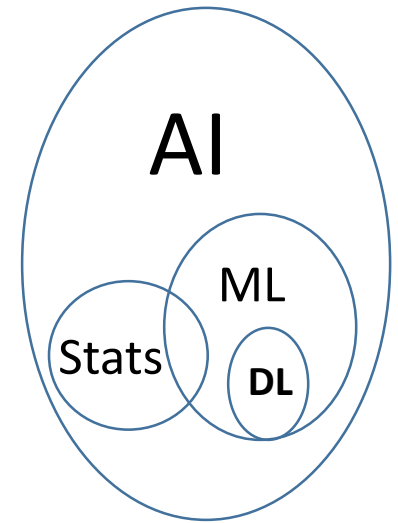
# Ingenuity Pathway Analysis Tool

**INGENUITY®**  
**PATHWAY ANALYSIS**

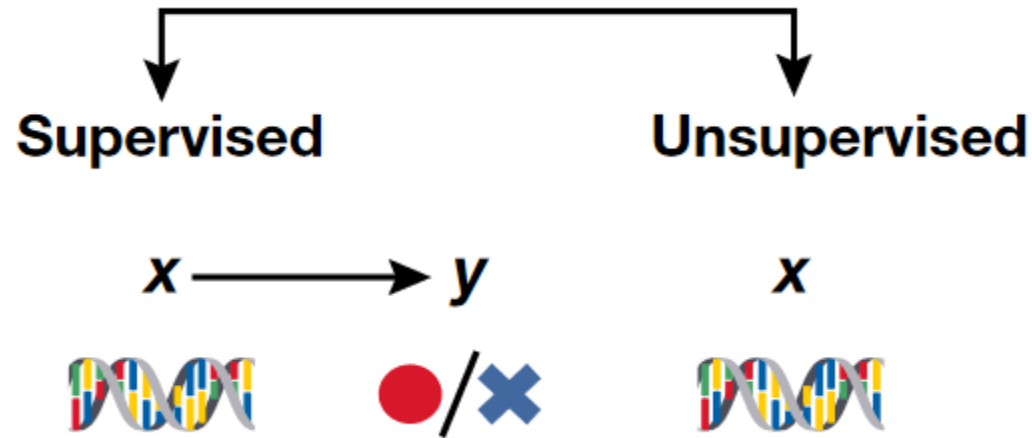


# Statistics vs. Machine Learning (ML)

- Statistics draws population inferences from a sample; ML finds generalizable predictive patterns
- Inferences vs Prediction. Models can be shared
- Statistics requires choosing model to incorporate our knowledge of the system; ML requires choosing a predictive algorithm by its empirical capabilities
- ML has limited applications in omics analysis



# Machine Learning Methods



- *Regression*
  - *Linear regression*
- *Classification*
  - *Logistic regression*
  - Random forest
  - SVM
  - Decision trees
- ...

- Dimension reduction (PCA, tSNE, UMAP, NMF, etc)
- Clustering (K means, hierarchical, etc)
- Factor analysis
- Outlier detection
- ...

- **Deep learning (DL) with neural networks**

*Generative vs.*

*Discriminative*

# Challenges and Limitations

- Curse of dimensionality, imbalanced class sizes, overfitting, high noise
- Limited amount of ground truth labeled data in genomics for supervised learning
- Supervised learning model interpretation from biological perspective



# Summary

- AI/ML methods are increasingly important in large scale omics data analysis
- Without careful consideration, practical utility of supervised learning in basic research is limited
- Deep learning holds great promises for functional genomics and clinical diagnosis
- Successful application of ML methods requires close collaboration of domain experts and data scientists