Differential Gene Expression

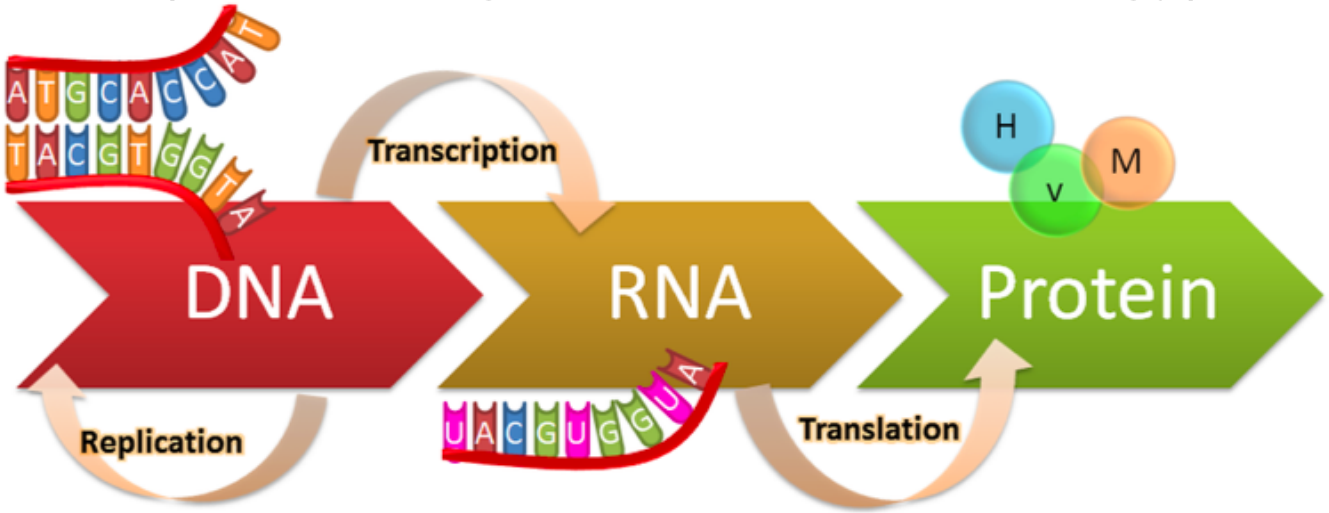
Differential Gene Expression

The “basic units” of DNA are called genes

Quantifiable “expression” levels

DNA → RNA → Protein

(Central dogma of molecular biology)



https://genius.com/Biology-genius-the-central-dogma-annotated

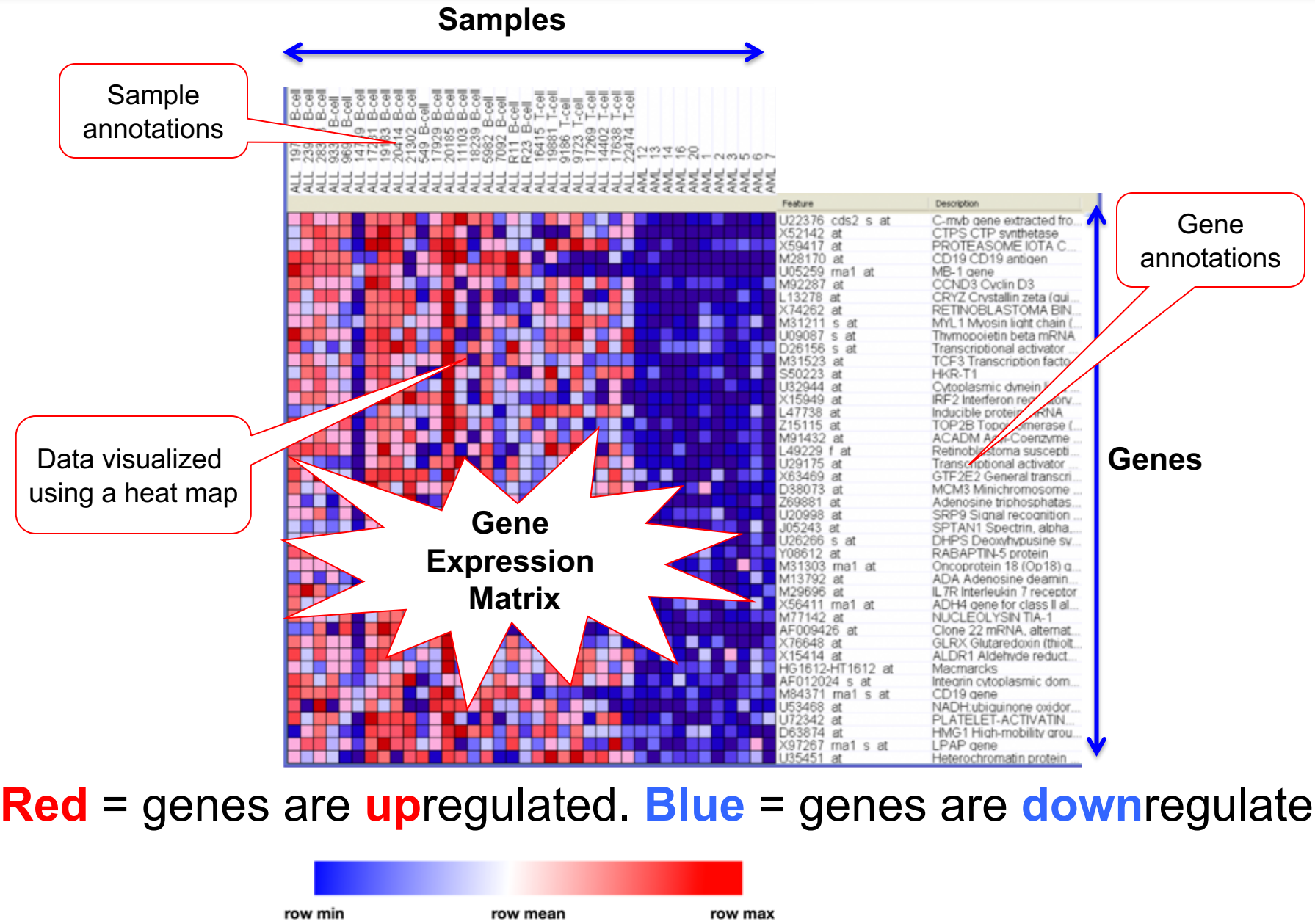
Active, ongoing process

Differential Gene Expression

Gene “A”, condition “1”
E.g., normal cell

Gene “A”, condition “2”
E.g., cancer cell

Data Representation and Visualization

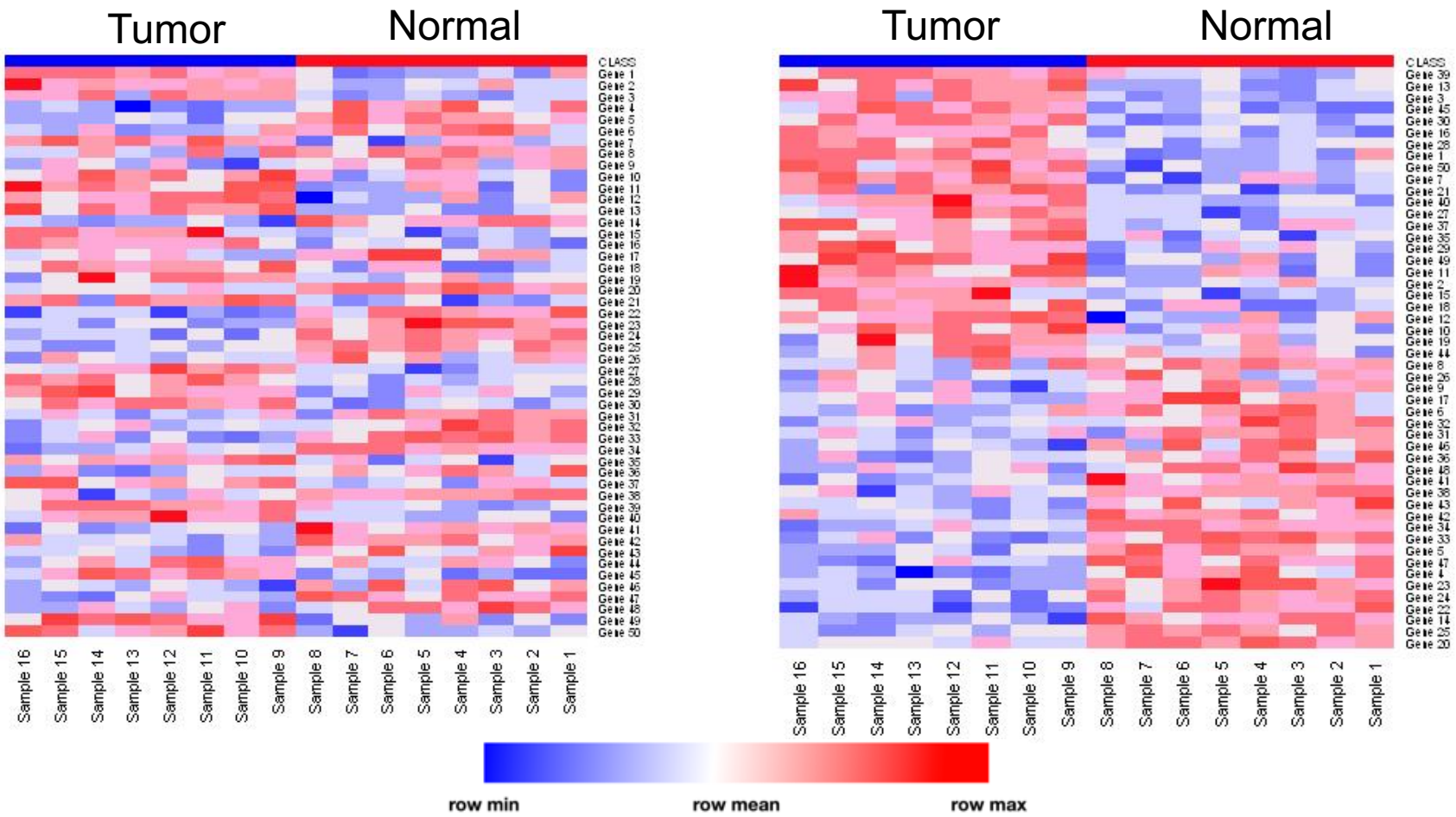


Red = genes are **up**regulated. **Blue** = genes are **down**regulated.

Goal of Differential Expression Analysis

Select "markers": given two distinct "phenotypes" (classes), find markers (genes) that distinguish these classes from one another

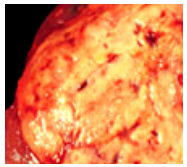
→ Note that markers work both ways (up and down regulated)



Gene Marker Selection

Hierarchy of difficulty

Problem	Gene Markers	Error	Example
I. Tissue or Cell Type Normal vs. Abnormal	~1000-2000	~0%	Normal vs. Renal carcinoma



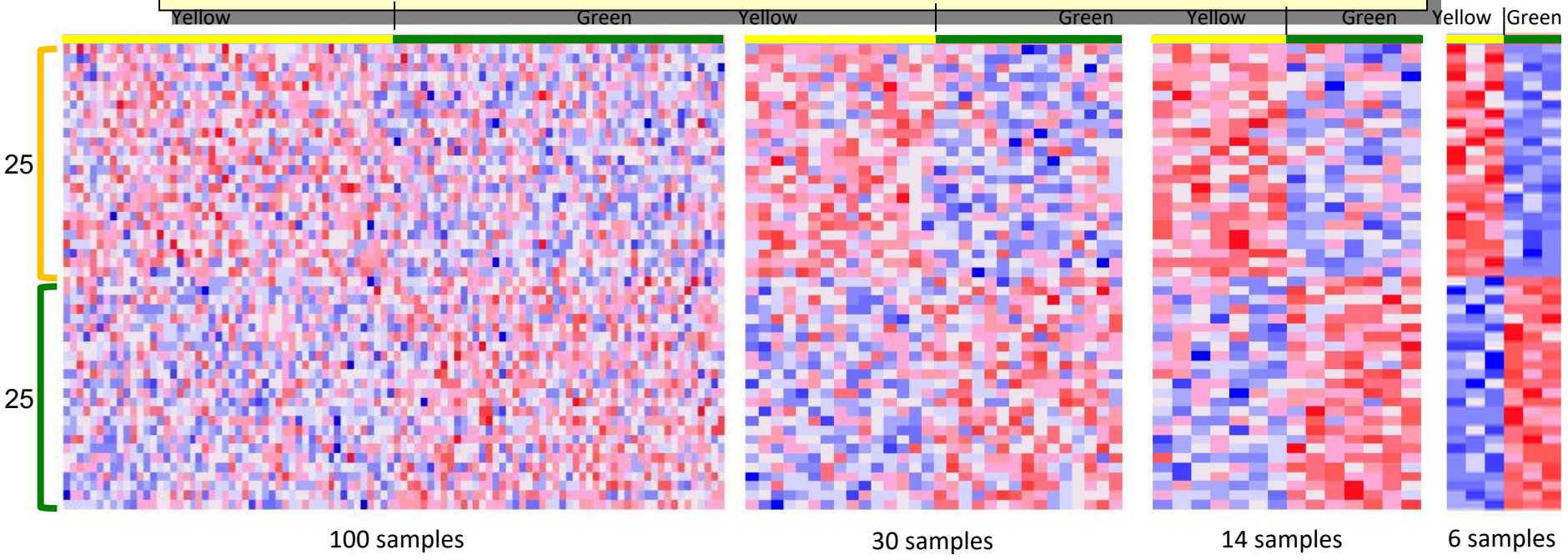
Degree of Difficulty

adapted from P. Tamayo

Effect of Sample Size

Exercise: select markers for random samples

- **Generate** a 10,000x100 matrix of **random data** → N(μ=0, σ=0.5)
- **Pick** *n* columns **at random** → n = [100, 30, 14, 6]
- **Assign** label yellow (e.g., tumor) to half of samples (chosen **at random**) and green (e.g., normal) the rest
- **Select** top 25 markers for yellow, top 25 markers for green



With small sample size it is easy to find genes considered significant by chance!

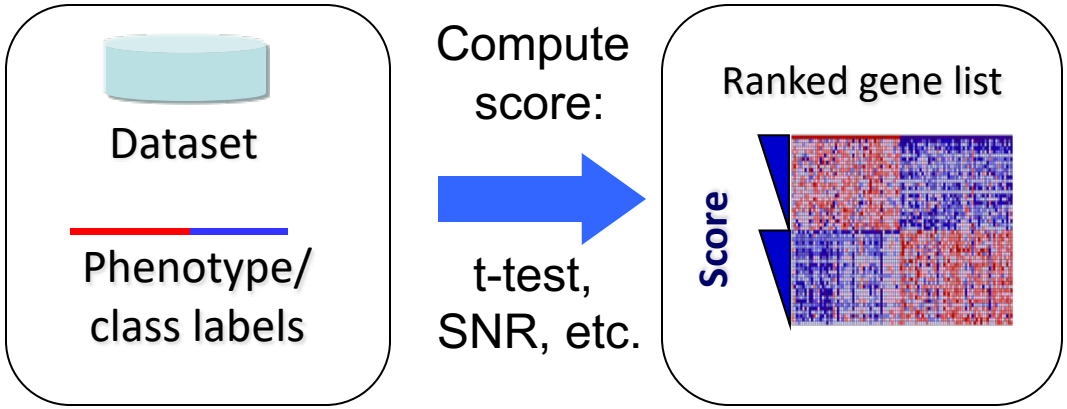
Differential Analysis Exercise

Open notebook:

2018-01-23_05_UBIC_Differential Analysis

Gene Marker Selection

Compute score for each gene



t-test $\frac{\mu_A - \mu_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$

Hypothesis testing method: Standardized mean difference between the two classes.

It is the difference between the mean expression of class A and class B divided by the variability of expression.

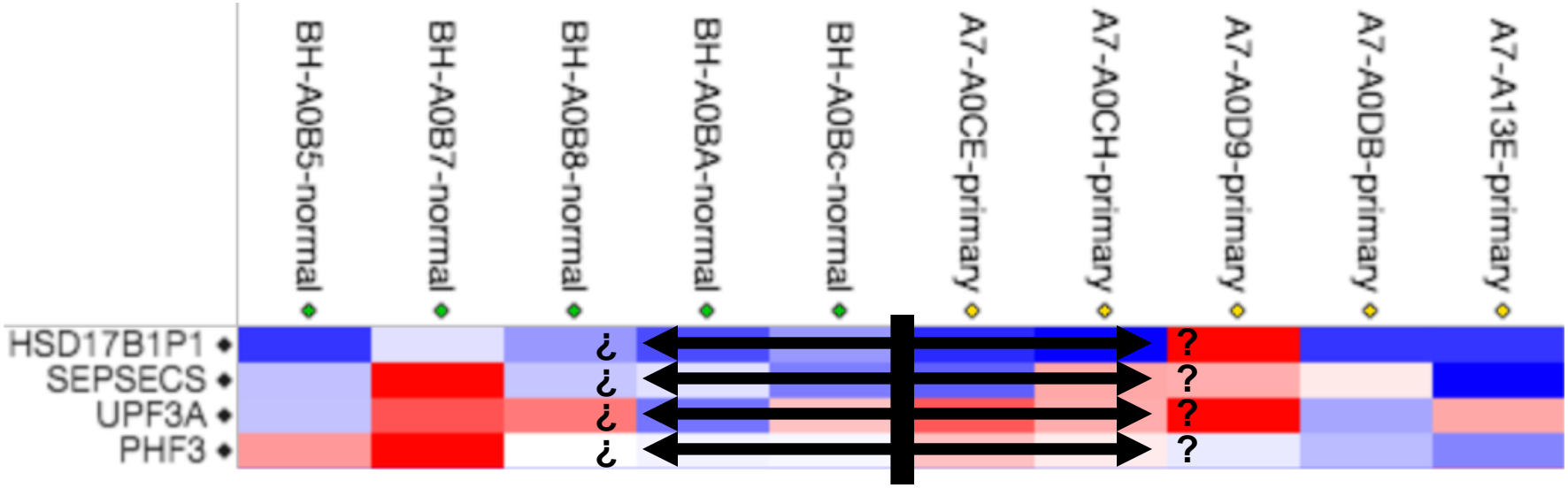
Signal-to-Noise Ratio (SNR) $\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$

Similar to the t-test but **takes the standard deviation of the two distributions into account** which is more representative of the differences between classes when there may be differences between the SD of class A and the SD of class B.

μ = class mean
σ = std deviation
n = # of samples

Multiple Hypothesis Testing

- Remember: Each gene/row is a hypothesis!



- Reduce number of hypotheses/genes by variation filtering (attempt at reducing false negatives)
- There are ways quantifying this such as False Discovery Rate