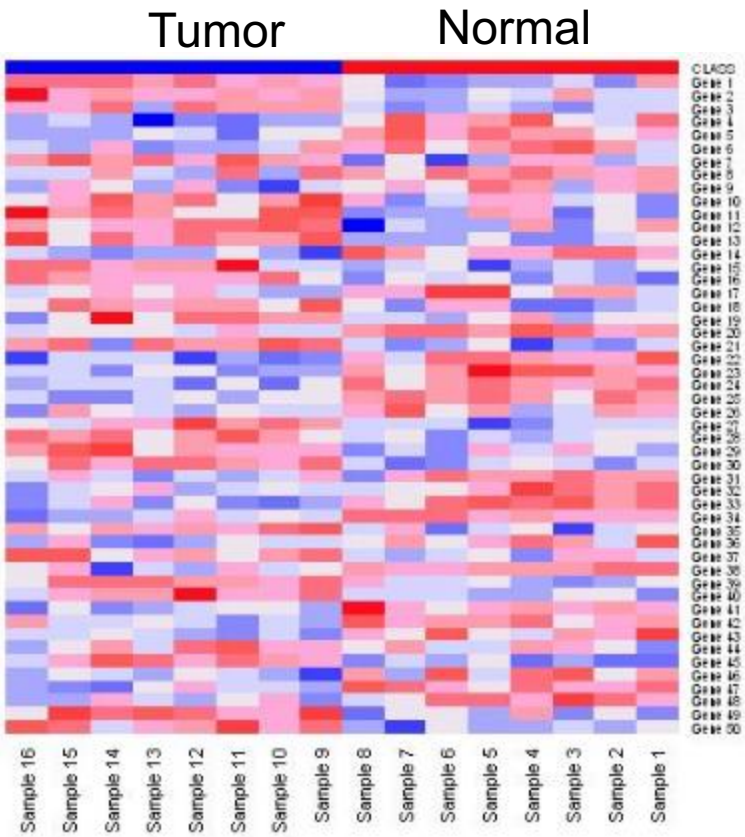
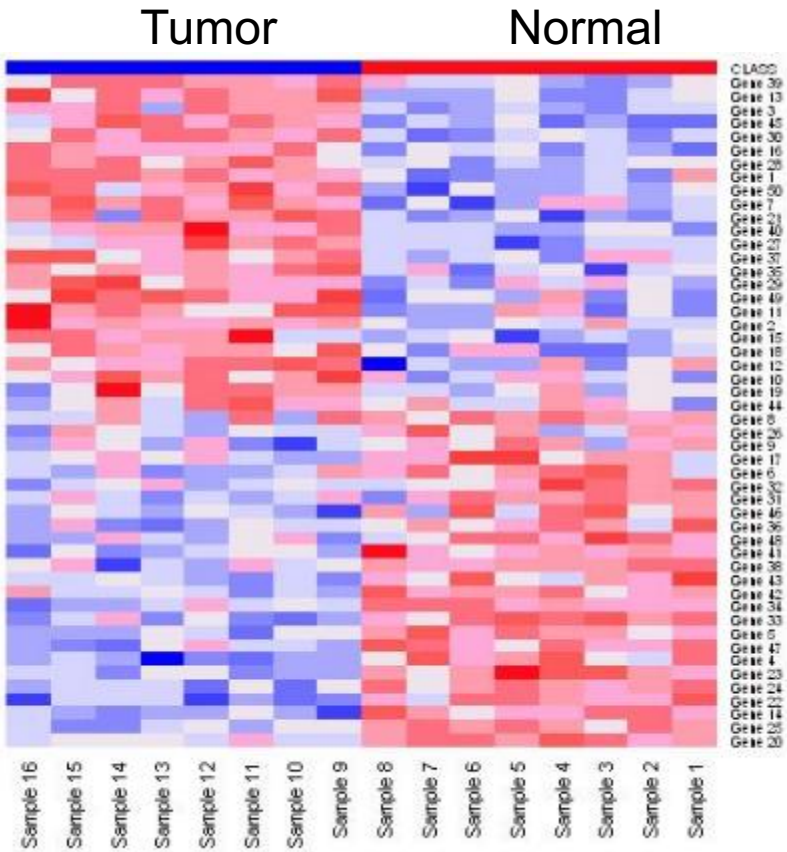
Differential Gene Expression

Differential Expression Analysis

Marker selection

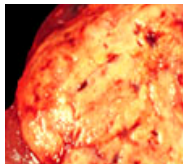
Given phenotypically distinct classes, find “markers” that distinguish these classes from one another





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  

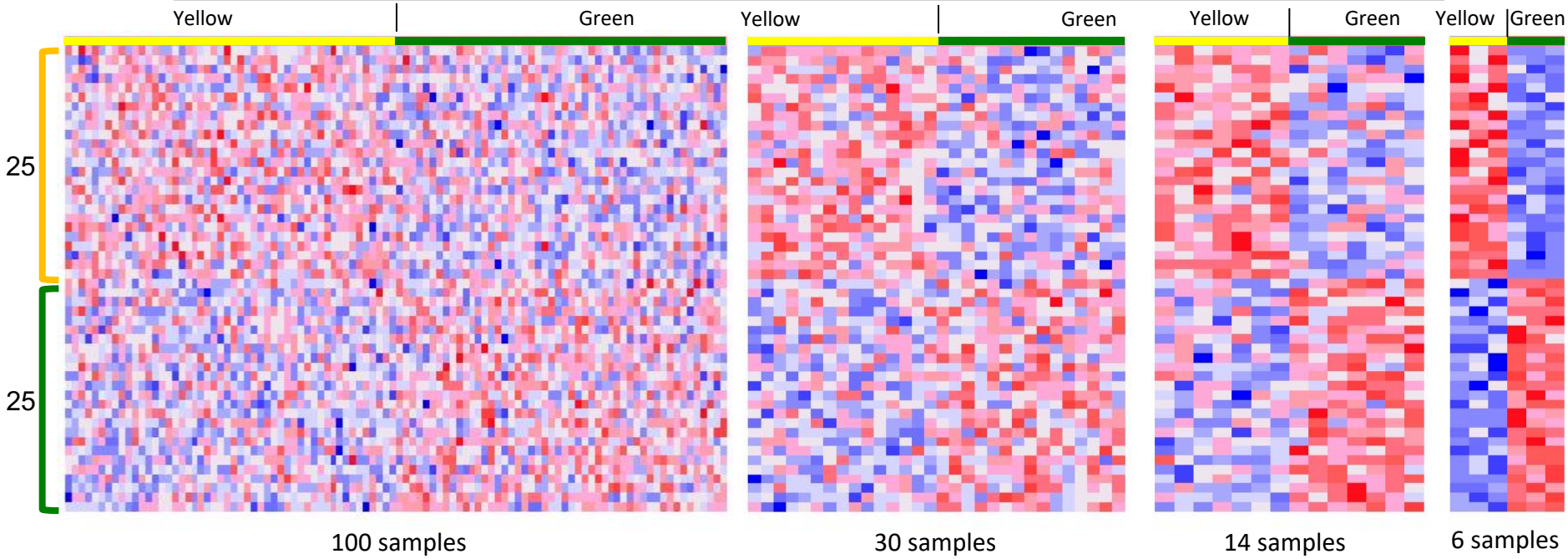
Degree of Difficulty

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 so few samples it is easy to find rows that look the way you want them to!

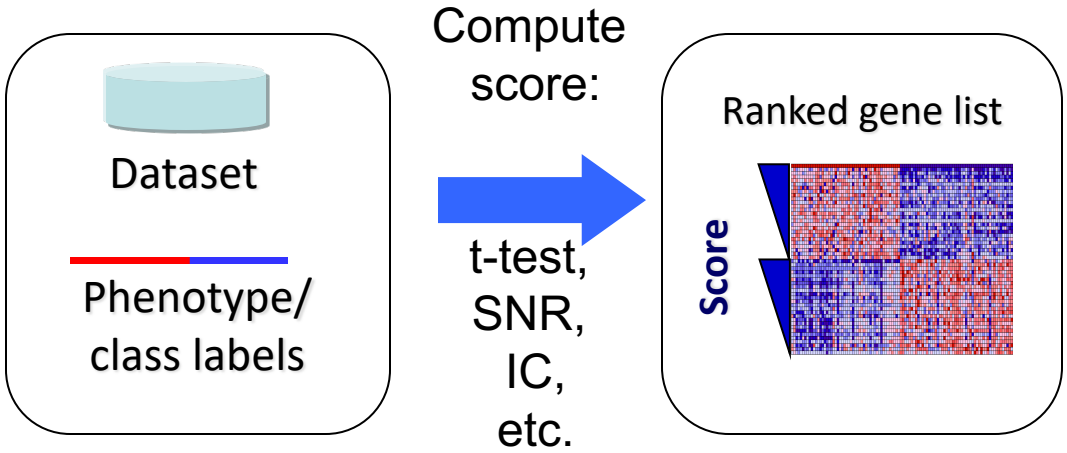
Differential Analysis Exercise

Open notebook:

2018-03-14_05_UCSF_Differential Analysis

Gene Marker Selection

Compute score for each gene



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

t-test

Hypothesis testing method:
It is the **difference between the mean expression** of class A and class B **divided by the variability of expression**.

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

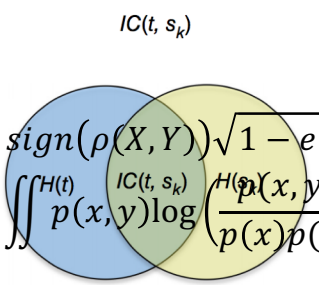
Signal-to-Noise Ratio (SNR)

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.

$$\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$$

Information coefficient (IC)

This test takes **the amount of shared information** between the two classes.

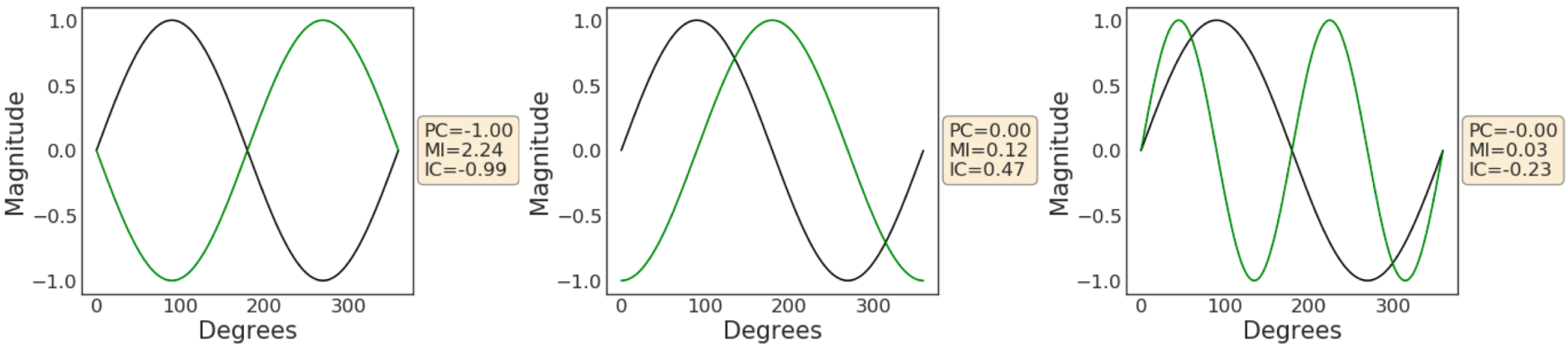


$$IC(X,Y) = sign(\rho(X,Y))\sqrt{1 - e^{-2MI(X,Y)}}$$

$$MI(X,Y) = \iint p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy$$

Mutual Information → Information Coefficient

- The Mutual information (MI) quantifies the **how well we can predict one variable if we know the other**
- MI is more sensitive than Pearson correlation (PC). Particularly, unlike PC, MI is **sensitive to nonlinear relationships** between variables
- The information coefficient (IC) is a normalized version of MI to keep its values between -1 and 1
- Computing the IC is computationally/time intensive



Differential Analysis Cookbook

1. Reduce number of hypotheses/genes by variation filtering (attempt at reducing false negatives)
2. If enough samples, compute p-values by permutation test (otherwise, compute asymptotic test using the standard t-distribution).
3. Control for Multiple Hypothesis Testing by using the FDR correction
   - Remember: if you choose FDR ≤ 0.05, you're willing to accept 5% of false positives.
   - If number of significant hypotheses/genes "too large" even for very small threshold values, either:
     - use the maxT correction (possible w/ empirical p-values only).
     - use additional criteria (e.g., min fold-change, min expression value, etc.)

Differential Analysis

GenePattern modules

- Create count data set – **download_from_gdc**
- Filter and transform data – **PreprocessReadCounts**
- Make class/phenotype file – **ClsFileCreator**
- Run Differential Analysis – **ComparativeMarkerSelection/DESeq2/OC Notebook**
  - Choose test statistic (say, Information Coefficient)
- View results with **ComparativeMarkerSelectionViewer**
  - If enough samples, compute p-values by permutation test (otherwise, use asymptotic test).
  - Control for Multiple Hypothesis Testing by using the FDR correction
  - Use **HeatMapViewer** to view results for top genes
- Use **GSEA** to find gene sets (or pathways) that are enriched in your dataset – coming up after the break!