Introduction to Systems Biology of Cancer

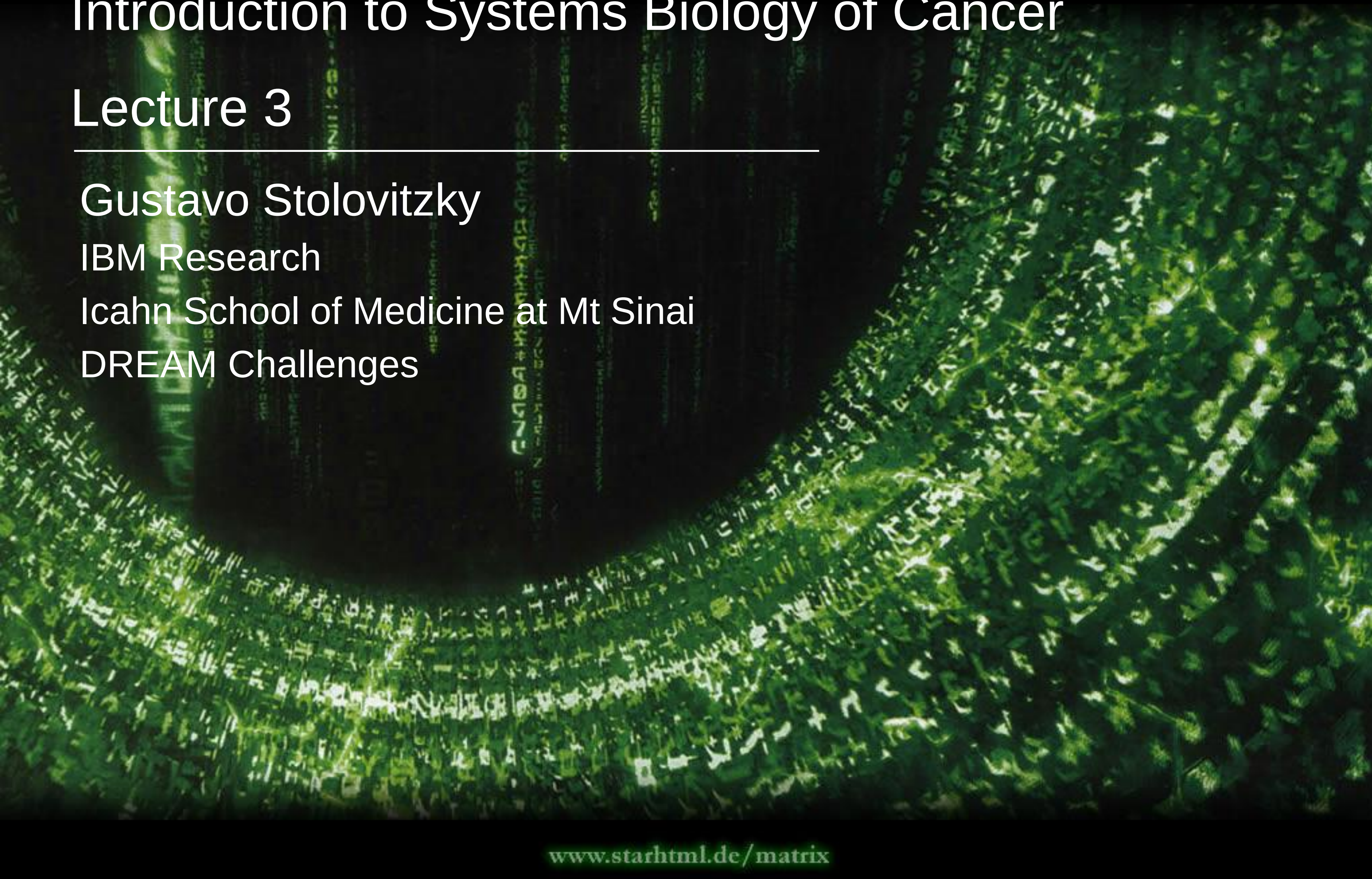
Lecture 3

Gustavo Stolovitzky

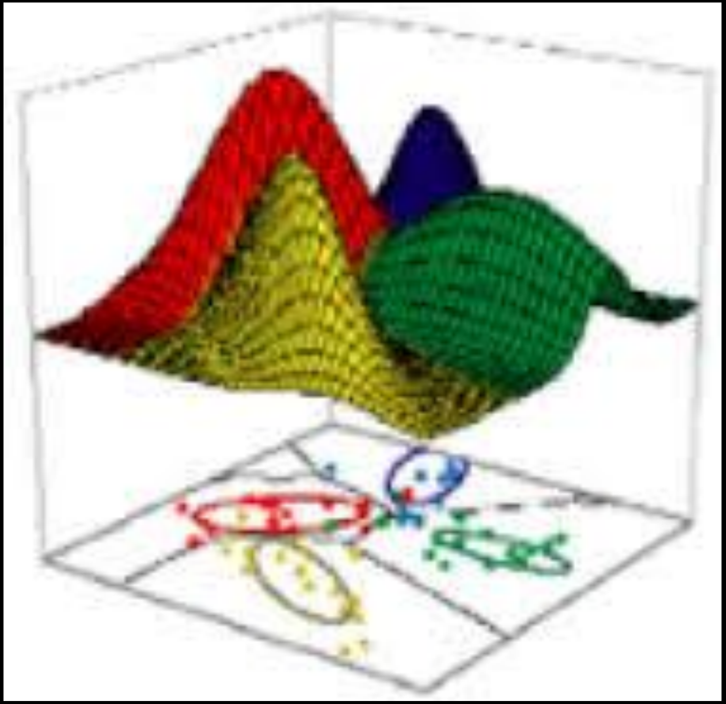
IBM Research

Icahn School of Medicine at Mt Sinai

DREAM Challenges



From MIT Course: Statstical Learning Theory and Applications



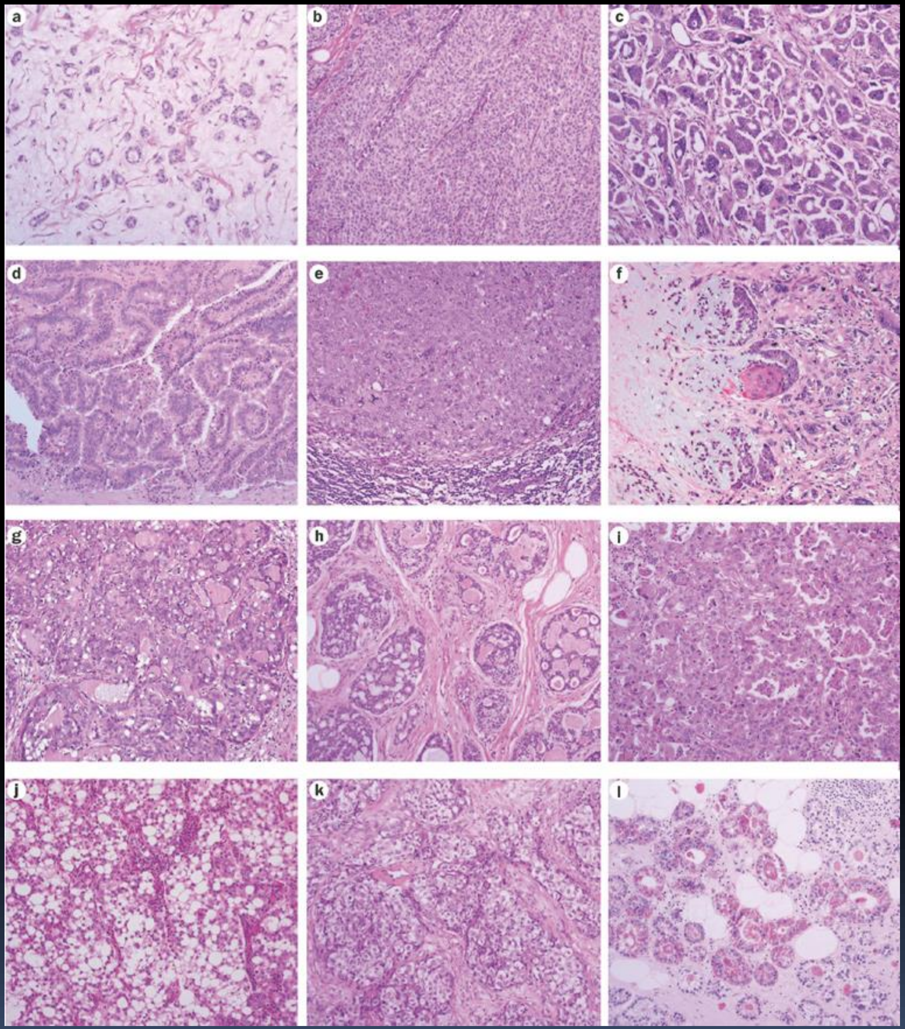
Classification of cancer

Traditional Classification of Cancer

- More than 200 types of cancer are commonly defined

- Clinicopathological Information further subdivides each cancer type
  - Demographic and Clinical history: gender, age, family history of cancer
  - Stage: size of tumor, lymph node involvement, presence of metastasis
  - Tumor specific: location, size, histology
- Pathologists take a thin slice of tumor (biopsy or surgery). Under the microscope they can assign the histological type and determine the grade and prognosis based on
  - Appearance of the cells
  - Size and shape of the nuclei
  - Differentiation of the tumor (how much the cell resemble normal cells)
  - Number of mitosis
  - Invasiveness

Histological Classification of Cancer



Histological special types of breast cancer

Nature Reviews Clinical Oncology
6, 718-730 (December 2009)

a | Mucinous carcinoma. b | Neuroendocrine carcinoma. c | Micropapillary carcinoma. d |Papillary carcinoma. e | Medullary carcinoma. f | Metaplastic carcinoma. g | Secretory carcinoma. h | Adenoid cystic carcinoma. i | Apocrine carcinoma. j | Lipid-rich carcinoma. k | Glycogen-rich carcinoma. l | Acinic cell carcinoma.

Traditional Classification of Cancer

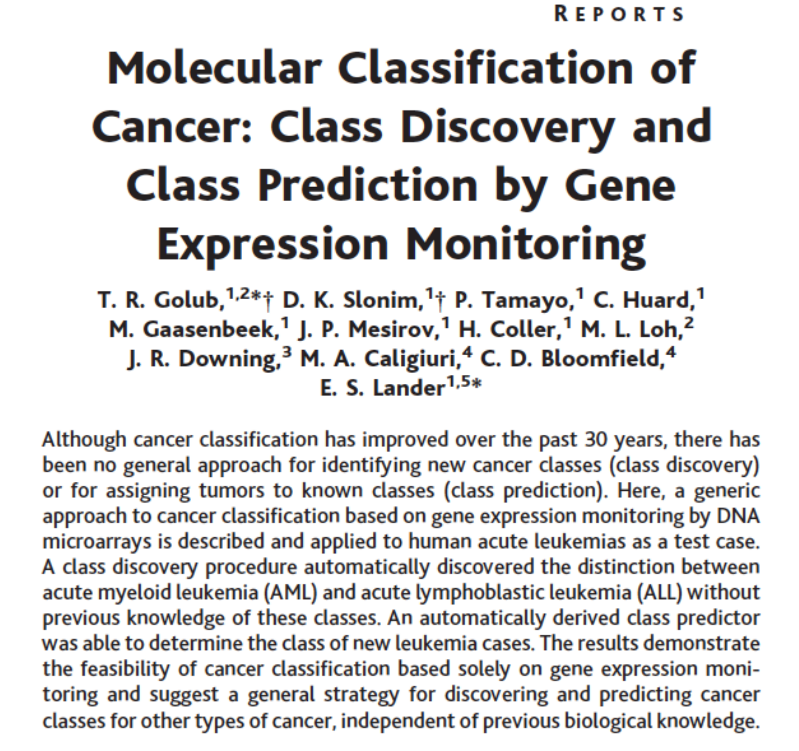
- These clinico-pathological parameters currently determine the therapy

- Some problems with this approach:
  - It depends on the histological section used
  - It depends on the pathologist:
    - In bladder cancer, a study showed that the concordance between pathologists in assigning grade/stage was of ~70%
    - This is worse for gliomas
  - Patients with the same clinicopathological parameters
    - Sometimes follow different clinical course.
    - Respond differently to therapy

- These problems suggest that we need a further classification

Molecular Classification of cancers

- The systematic profiling of various cancer types was amongst the first applications transcriptomics. The seminal paper in the field is
- Golub et al. for class discovery and prediction in AML and ALL

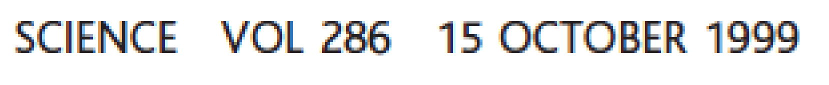


REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case. A class discovery procedure automatically discovered the distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. An automatically derived class predictor was able to determine the class of new leukemia cases. The results demonstrate the feasibility of cancer classification based solely on gene expression monitoring and suggest a general strategy for discovering and predicting cancer classes for other types of cancer, independent of previous biological knowledge.



SCIENCE VOL 286 15 OCTOBER 1999

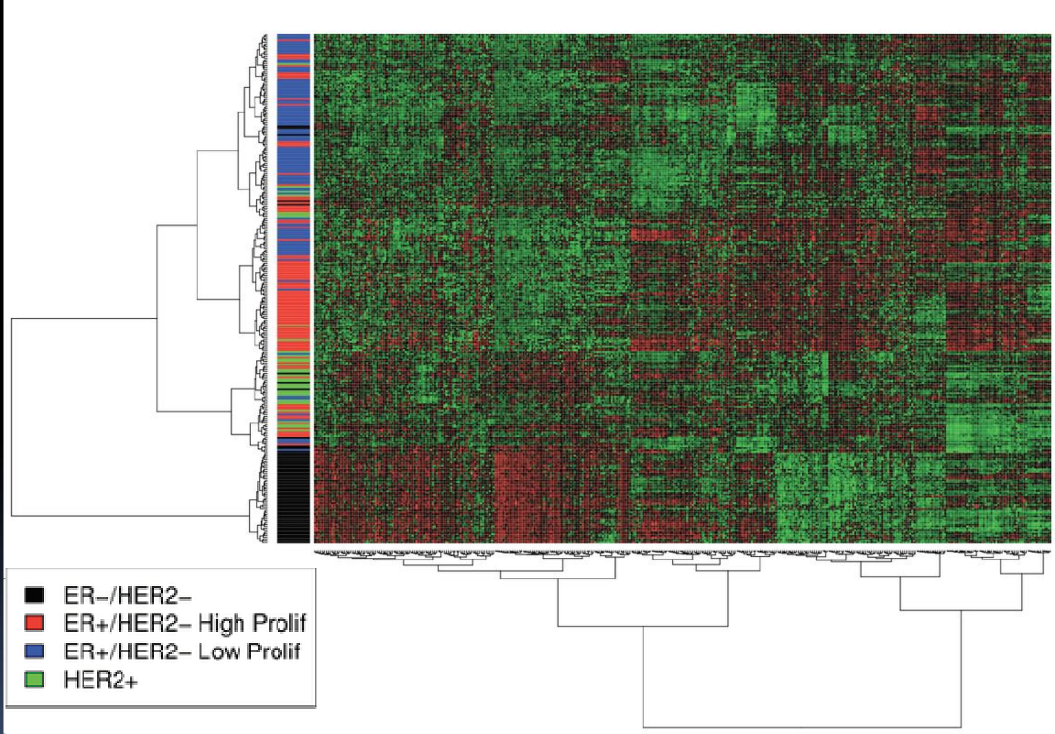
Example of molecular classification of Breast Cancer

- We aim to discover homogeneous subtypes within a collection of tumors (data from www.thelancet.com Vol 365, pag 671, 2005)

p = 12,065 genes

(reduced to 400)

n = 286 patients



- Notice that the histology characterization coincides very closely with the molecular-based grouping.

How was the grouping done?

- First they reduced the number of features to 400 by only using the ones that have the highest variance across the 286 patients.
- The genes and patients were reordered so that they show the same properties according to their expression in the two dimensions. This makes the visualization more intuitive.
- A dendrogram allows us to visualize the hierarchical tree like structure of the data.
- This way to visualize a large data matrix is called a heatmap and was popularized in comp bio by Michael Eisen.

What does the grouping tell us?

- The classical subtypes based on biomarkers and mitosis (ER, HER2, and proliferation) are largely recovered (but not completely) if we cut the dendrogram at a depth corresponding to 4 clusters.

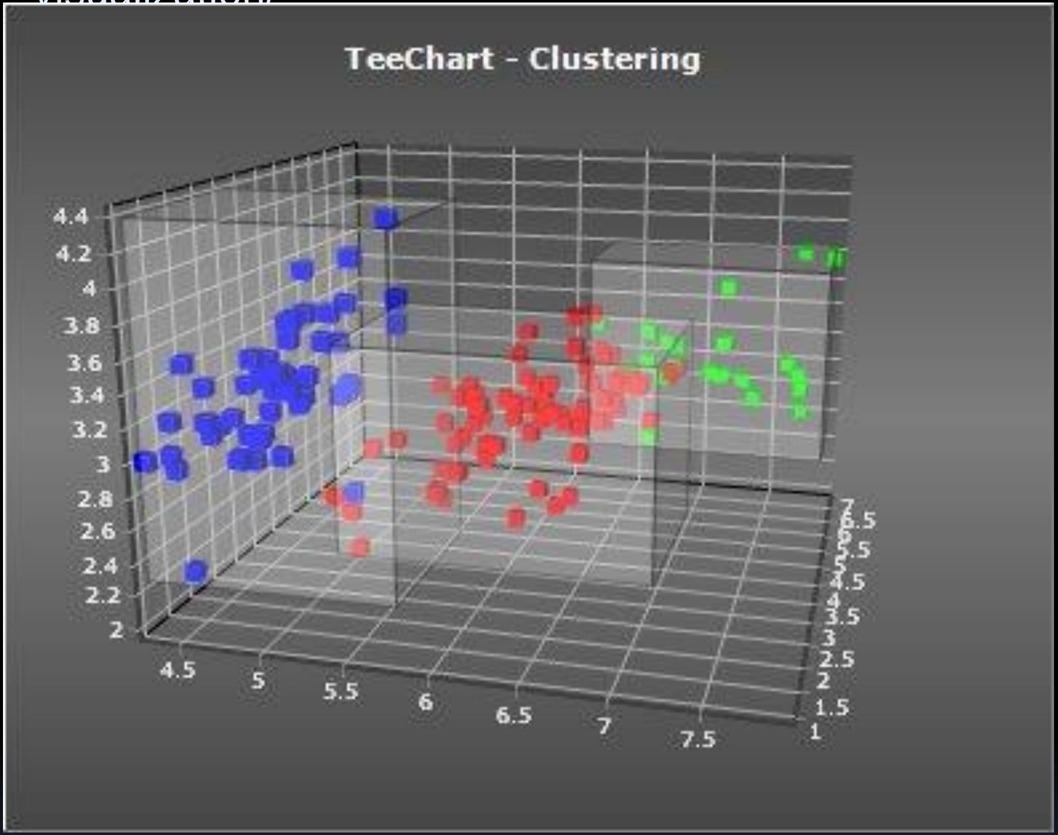
- This suggests that a automatic and biologically relevant classification of cancers from omics is possible.

- Let us focus on the algorithms for grouping. The ones that we just showed are called clustering or unsupervised classification.

- There is a universe of clustering methods. Next will just see a few.

Clustering

From https://steema.com/wp/blog/2015/06/01/clustering-visualization/



Clustering

- Let *X* be and *n* x *p* matrix, with *p* genes measured in *n* samples

- Distance: Clustering requires a notion of similarity or distance. If we want to group samples into a small number *k* << *n*, we need that the elements within a group (cluster) be more similar than elements of different groups. Popular distances are the l_q distance

$$||X_{i*} - X_{j*}|| = \left(\sum_{k=1}^{p} |X_{ik} - X_{jk}|^q \right)^{1/q}$$

- l_2 is clearly the Euclidian distance, and l_1 is the Manhattan distance. Or the Pearson correlation similarity

Clustering

- If the data needs to be normalized, a Pearson correlation is a good choice

$$r(X_{i*}, X_{j*}) = \frac{\sum_{k=1}^{p}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_{k=1}^{p}(X_{ik} - \bar{X}_i)^2 \sum_{k=1}^{p}(X_{jk} - \bar{X}_j)^2}}$$

where

$$\bar{X}_i = \frac{1}{p}\sum_{k=1}^{p} X_{ik}$$

Pearson is a similarity coefficient. It can be transformed into a distance by the operation 1-*r.* When the mutual relation between two samples is non-linear, other choices may be more appropriate, such as the Spearman correlation or the Mutual Information.

Hierarchical clustering

- Several algorithms exist.
  - Agglomerative: bottom up clustering
  - Divisive: when groups are divided in a top down strategy.

- Linkage function: how the distance between clusters of patients are computed. Given two groups of patients A and B, we have
  - Average Linkage

$$L(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

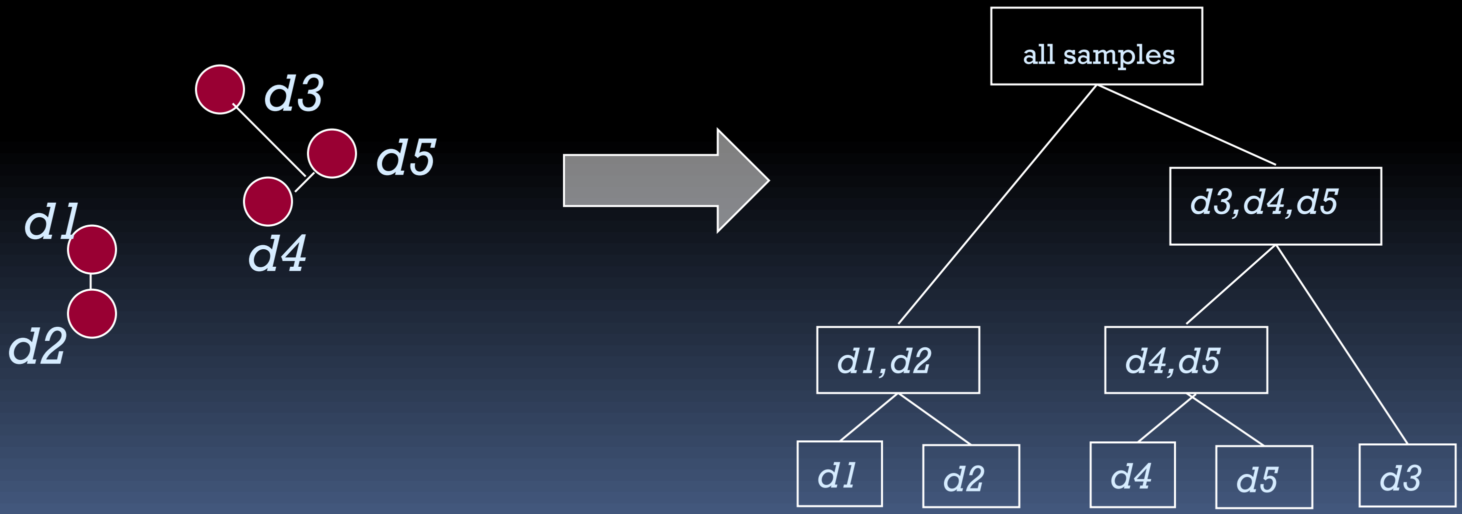
  - Centroid Linkage

$$L(A,B) = d\left(\frac{1}{|A|} \sum_{a \in A} a, \frac{1}{|B|} \sum_{b \in B} b \right)$$

Hierarchical agglomerative clustering

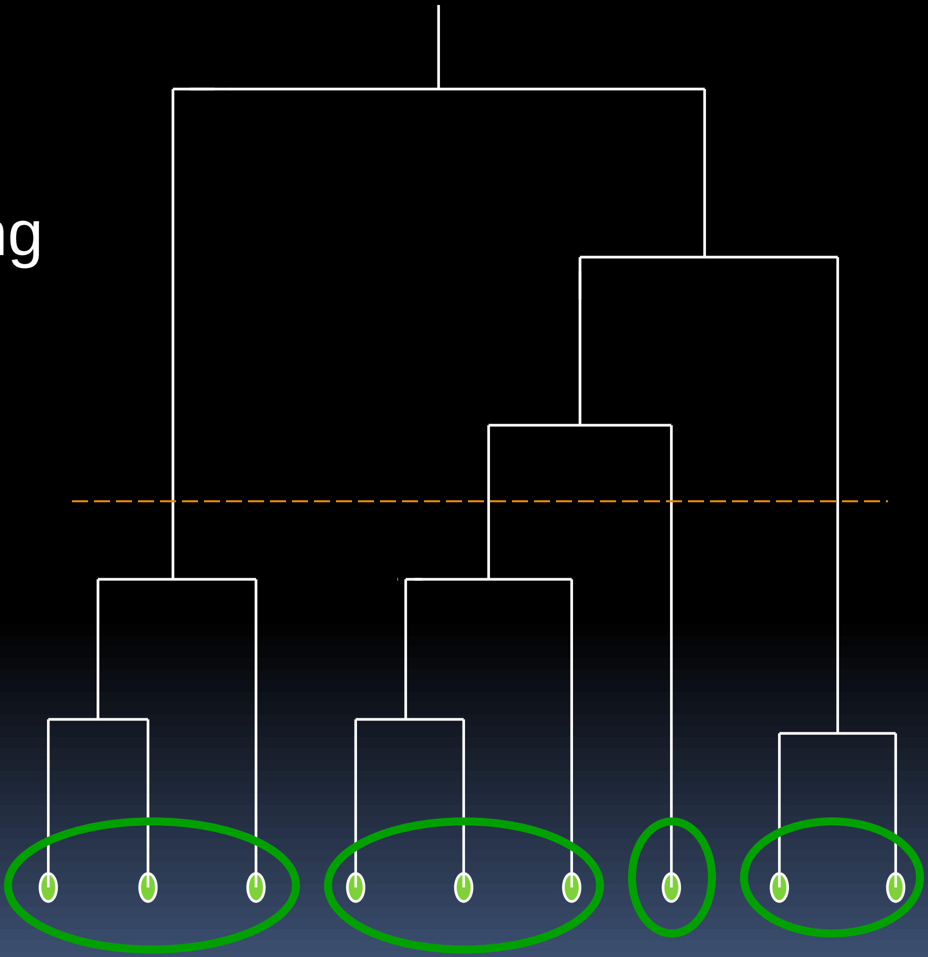
- Algorithm for agglomerative clustering

```
Start with all instances in their own cluster.
Until there is only one cluster:
    Among the current clusters, determine the two
        clusters, $c_i$ and $c_j$, that are most similar.
    Replace $c_i$ and $c_j$ with a single cluster $c_i \cup c_j$
```



Dendrograms

- At the end of the process clusters are obtained by cutting the dendrogram at a desired level
- each connected component forms a cluster.



Partitioning Algorithms

Goal: Construct a partition of a dataset *D* of *n* patients into a set of *k* clusters

Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion

- *Global optimal*: exhaustively enumerate all partitions (impractical)

$$E(K)=\Sigma_{j=1}^{K}\Sigma_{\mathbf{x}\in C_j} d^2(\mathbf{x},\mathbf{m}_j)$$

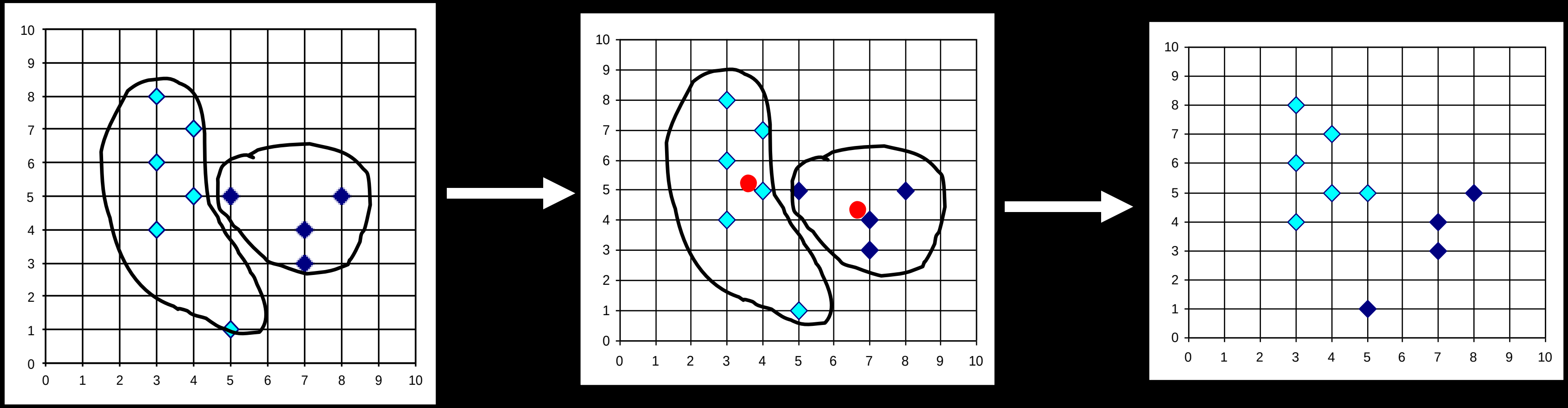
- Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen’67): Each cluster is represented by the centroid of the cluster
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw’87): Each cluster is represented by one of the objects in the cluster

K-means algorithm

Given *K*, the *K-means* algorithm is implemented in 4 steps:

- 1. Randomly assign objects into *k* nonempty subsets
- 2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
- 3. Assign each object to the cluster with the nearest seed point.
- 4. Go back to Step 2, stop when no more new assignment.

K-means algorithm: simple example

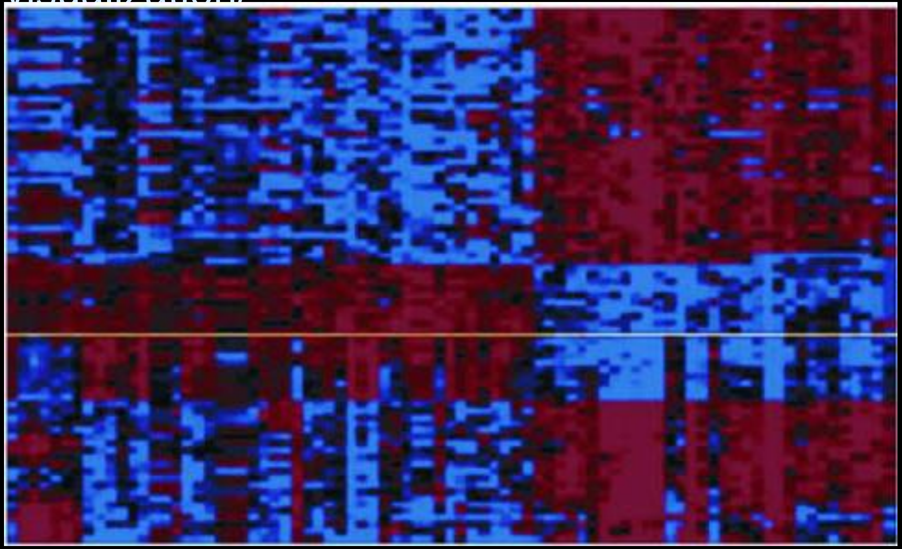


- Often terminates at a local optimum. Run many times and choose the one that gives the minimum of the cost function

$$E(K) = S_{j=1}^{K} S_{\mathbf{x}\hat{\imath}\ C_j} d^2(\mathbf{x}, \mathbf{m}_j)$$

- Need to specify K, the number of clusters, in advance. Chose the K at the “elbow” of E(K) vs K.
- Trouble with noisy data and outliers
- Not suitable to discover clusters with non-convex shapes

From https://steema.com/wp/blog/2015/06/01/clustering-visualization/



Identifying Differential Expression

Differential Expression Analysis

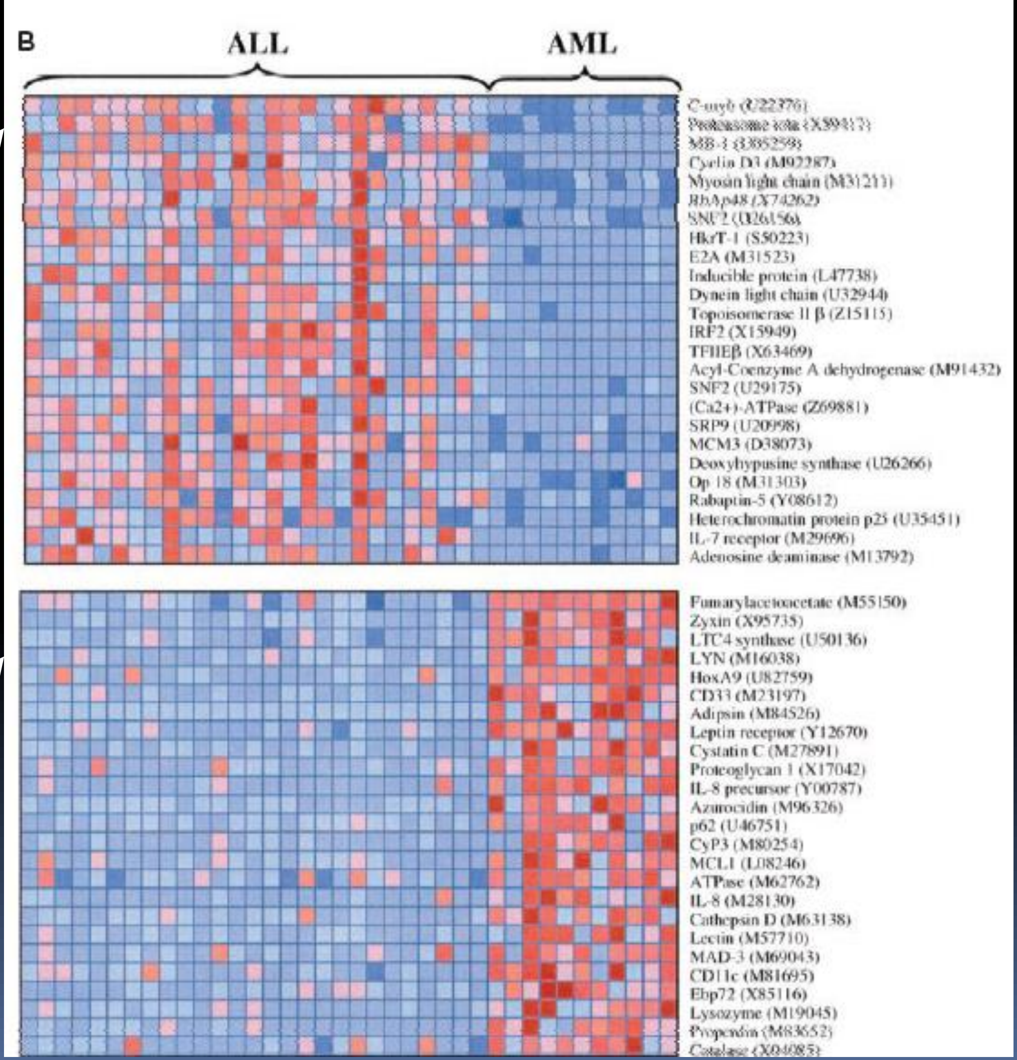
- This area of Systems Biology aims to answer the following question:
  - Given two conditions (Treated vs Untreated, Cancer vs Control, etc.), which are the genes that are expressed more in one condition than in the other?
  - Is this difference statistically significant?
  - Many classic statistical tests are available

Uses of Differential Expression Queries

- To find genes that are markers of health/disease status/progression
- To find genes that are markers of certain phenotypes
- To find the pathways that are specific to a phenotype
- To find the genes that respond to a drug or other perturbations
- To find genes that change in time t vs. time t_0

Classifying leukemia (Golub et al 1999)

class labels: 111111111111111111111111111 00000000000



genes upregulated in ALL compared to AML

genes upregulated in AML compared to ALL

Identifying differential expressed genes

Welch-t test

Assume $X_1, \ldots, X_m$ are gene expression values for a given gene in condition 2 and $Y_1, \ldots, Y_n$ correspond condition 2.

We compute

$$\bar{X} = \sum_{i=1}^{m} X_i /m \qquad \bar{Y} = \sum_{i=1}^{n} Y_i /n$$

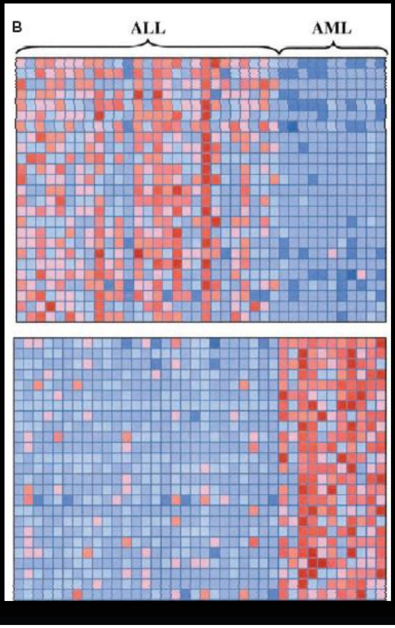
$$\sigma_1^2 = \sum_{i=1}^{m} (X_i - \bar{X})^2 /m \qquad \sigma_2^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 /n$$

and define the *t* statistics as

$$t = \frac{\bar{X} - \bar{Y}}{(\sigma_1^2/m + \sigma_2^2/n)^{1/2}}$$

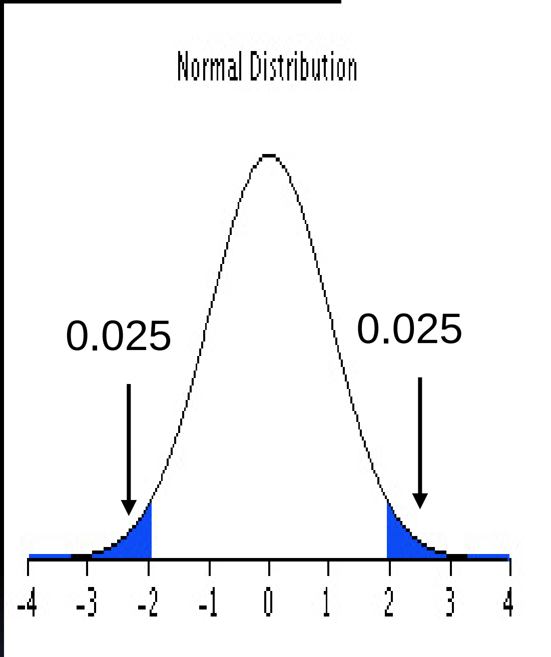
Identifying differential expressed genes

Welch-t test

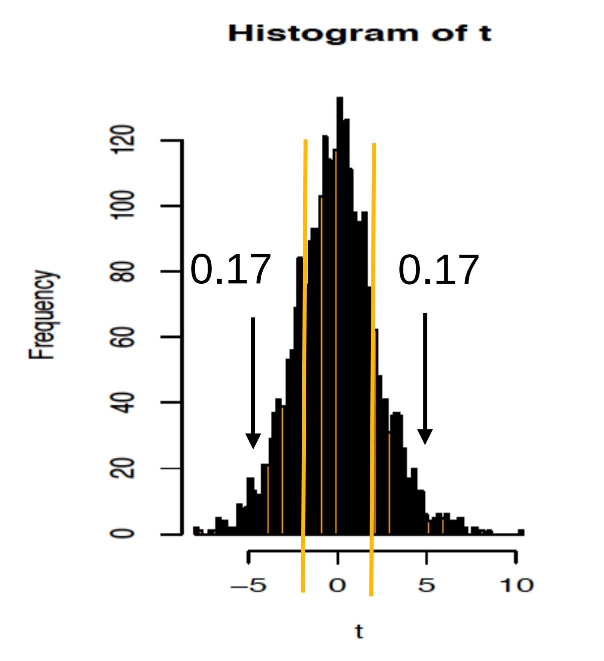


27 ALL vs. 11 AML samples, 3,051 genes.

Null Hypothesis



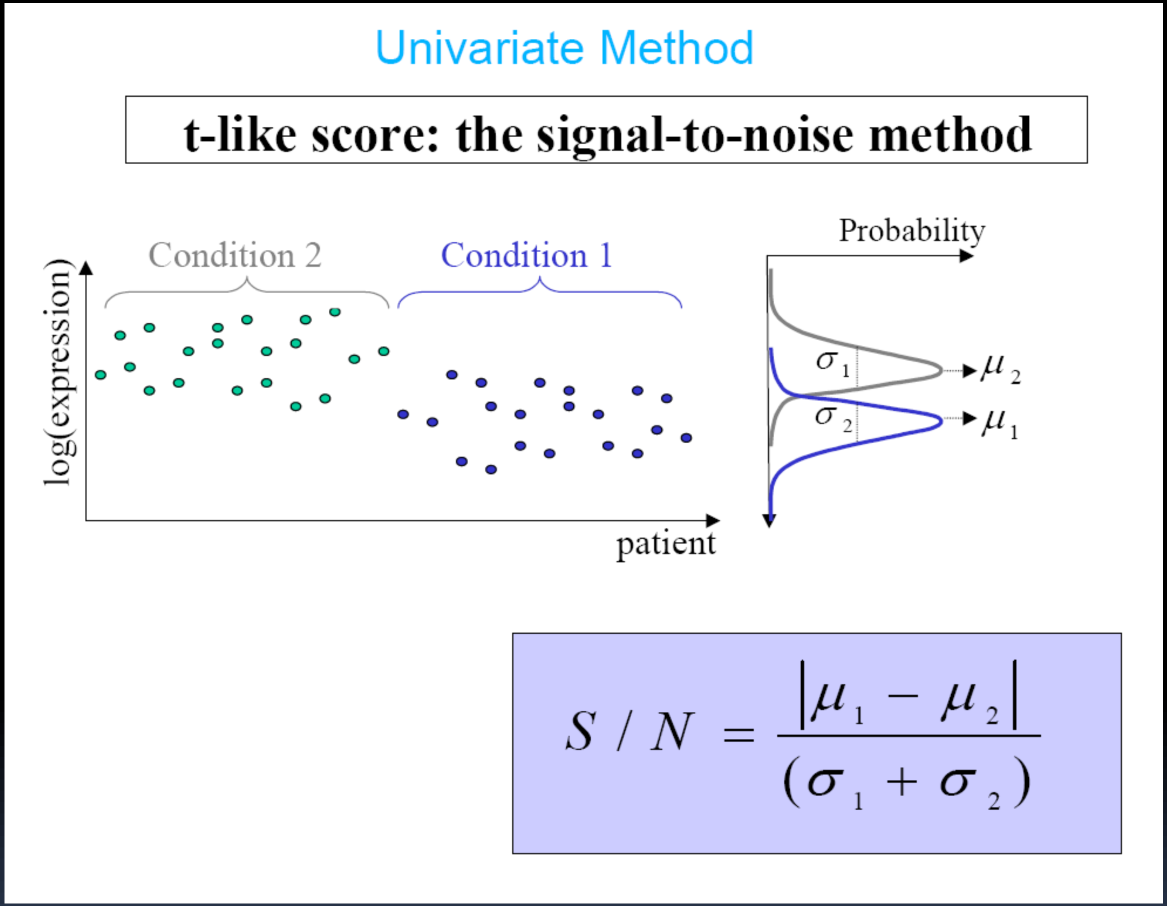
$$t = \frac{\bar{X} - \bar{Y}}{(\sigma_1^2/m + \sigma_2^2/n)^{1/2}}$$



If there were no effect (i.e., the means are the same), there should be a 5% of genes that have |t| > 1.96. Instead, we have a proportion of 1045/3052 = 34% >> 5%.

Our FDR is 5/34=~15%

Identifying differential expressed genes



Univariate Method

t-like score: the signal-to-noise method

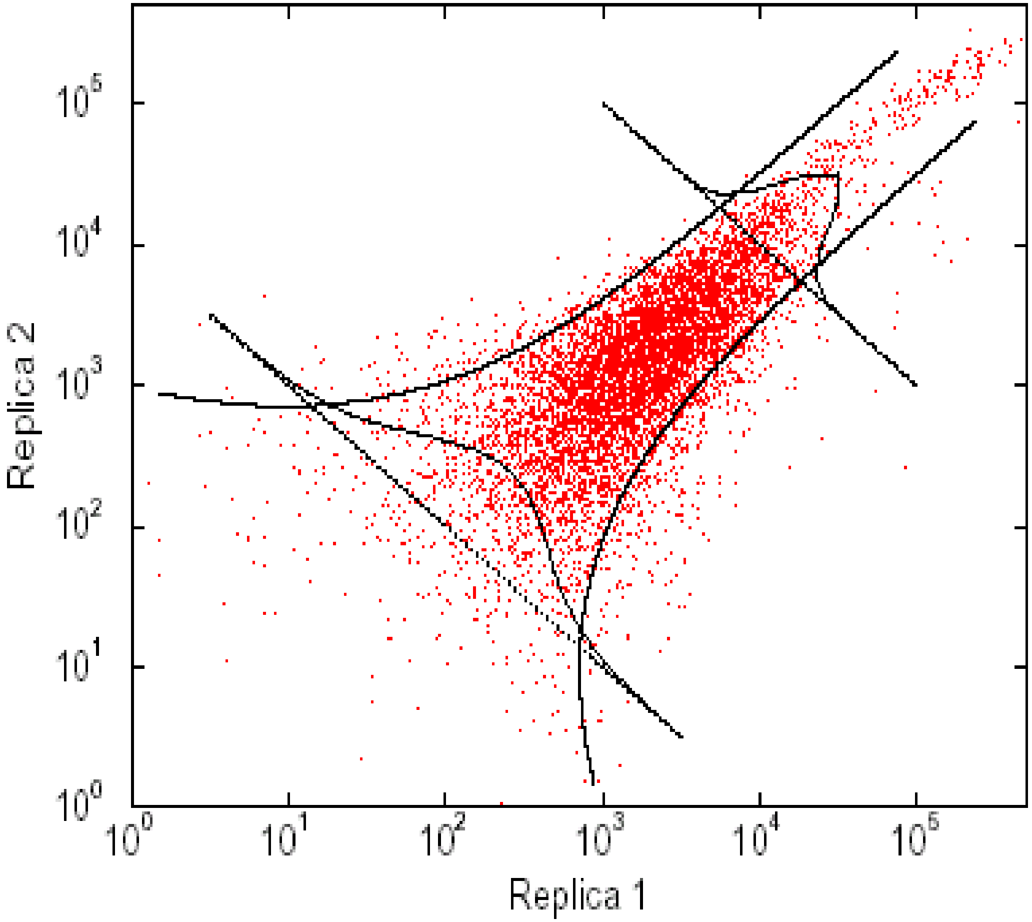
$$S / N = \frac{|\mu_1 - \mu_2|}{(\sigma_1 + \sigma_2)}$$

A statistical test needs to be performed to determine if the value obtained for a given gene has a signal to noise ration bigger than expected by chance.

Univariate Method

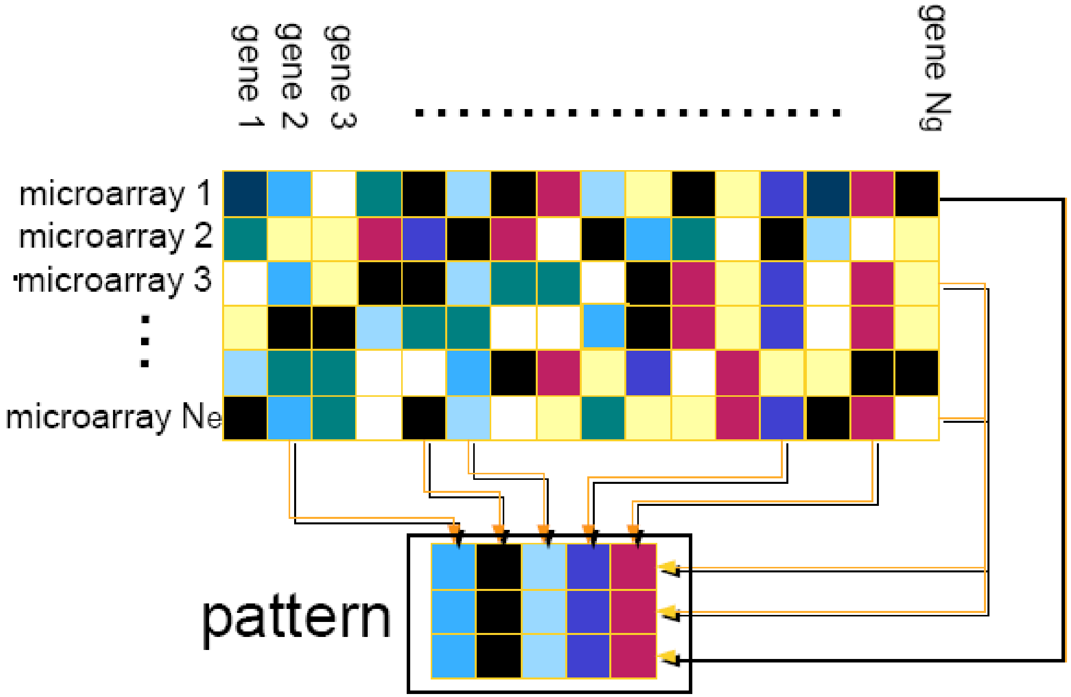
Noise based method: the USE-fold method

Good when we don't have replicas



Genes@Work

Multivariate Method



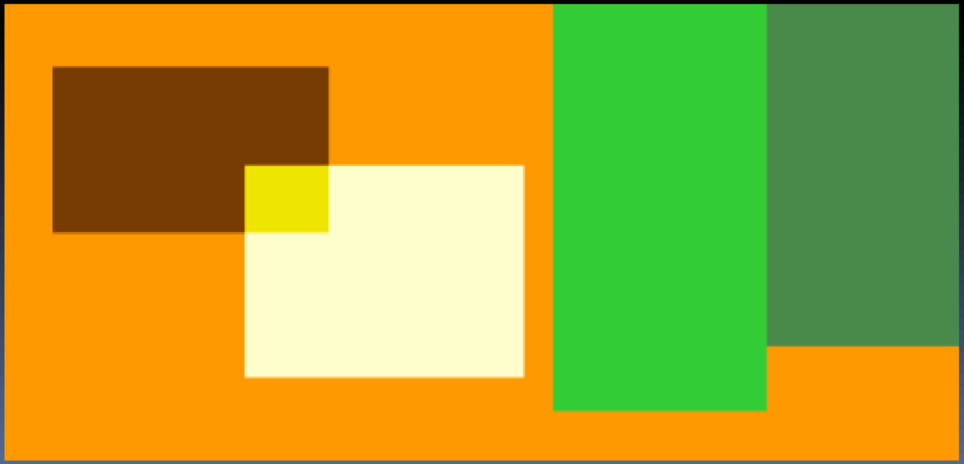
$$N_{jk}(j,k,N_e,N_g,\delta) \sim \binom{N_g}{k}\binom{N_e}{j}\alpha^k(1-\alpha)^{N_g-k}[1-(1+j^{-1})^k\delta^k]^{N_e-j} \leftarrow \alpha = j\delta^{j-1}-(j-1)\delta^j$$

$$p = 1-\exp\{-N_{jk}\}$$

Califano A, Stolovitzky G, Tu Y: "Analysis of gene expression microarrays for phenotype classification." *Proc Int Conf Intell Syst Mol Biol* 2000, 8:75-85.

Genes@Work belongs to a class of methods called “biclustering”

- These algorithms find statistical signal from the patterns (clusters) that are discovered in the data.
- These algorithms identify genes w/ common pattern across a subset of conditions
- The problem: Given an n x m matrix, A, find a set of submatrices, Bk, that satisfy some specific requirement that depend on the problem.



Each methods emphasize a different set of genes

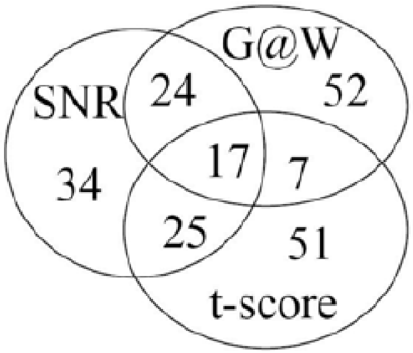
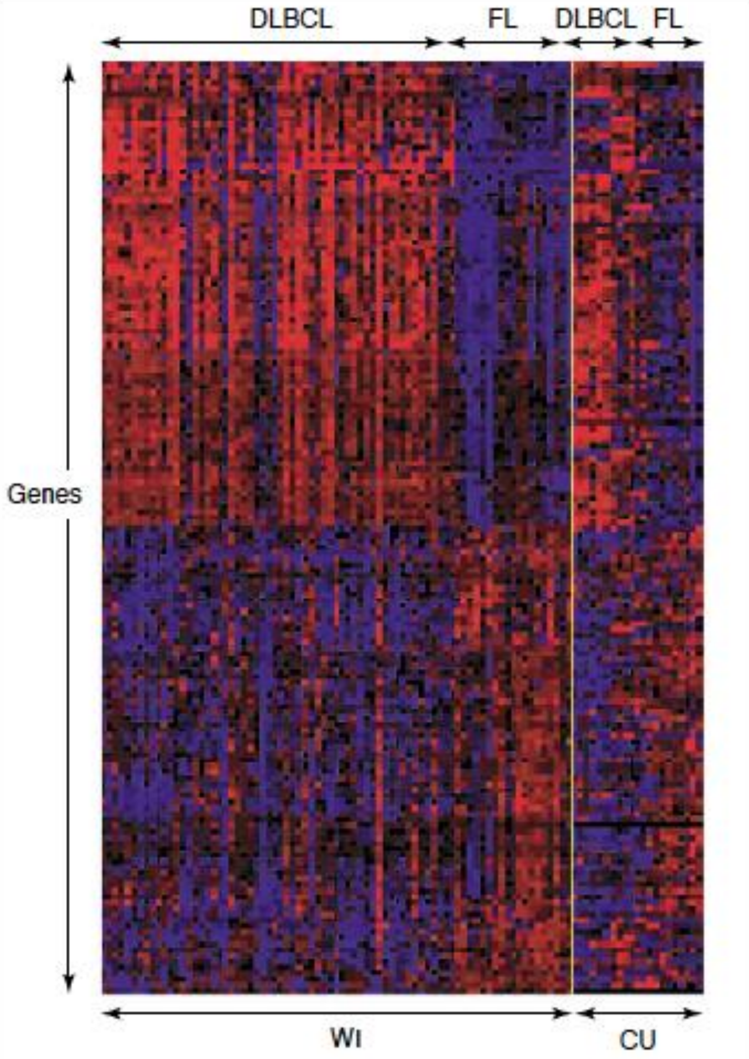


Figure 1. Venn diagrams of the set of genes identified in the analysis of diffuse large B-cell lymphoma and follicular lymphoma by each of three methods: the signal-to-noise ratio (SNR), Genes@Work (G@W), and the *t*-score. The numbers indicate the number of genes in each of the sets.



Current Opinion in Structural Biology

How do we know that our results are not due to chance?

- Statisticians developed a set of methods called hypothesis tests. In a nutshell, I want to see if the signal to noise statistics

$$\mathrm{SNR}_i = \frac{\mu_{i(1)} - \mu_{i(2)}}{\sigma_{i(1)} + \sigma_{i(2)}},$$

is considerably bigger when I use the right class labels

True class labels: 111111111111111111111111111 00000000000

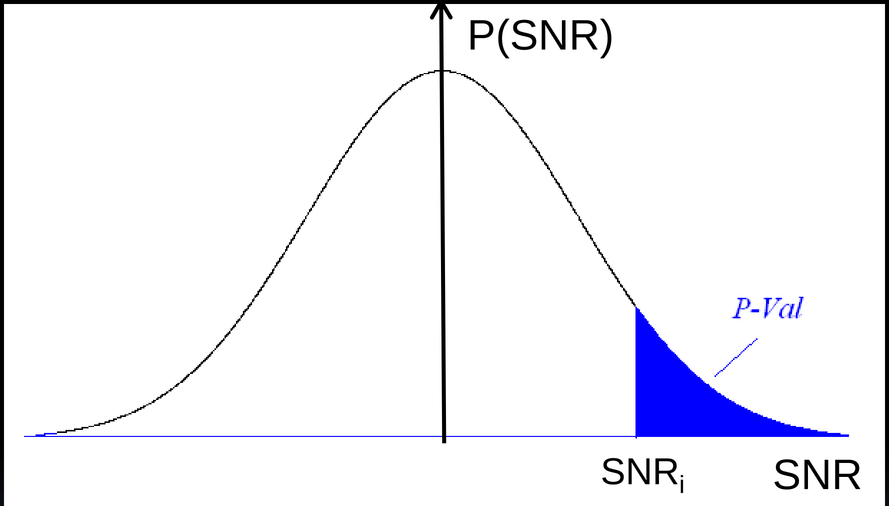
Compared to the situation in which the class labels are randomized

Randomized class labels: 101110111110111101111101101 1010100101010

The assumption that the class labels are random is called the Null Hypothesis. I create an ensemble with many permutations of the class labels, and for each I compute a measure of the “signal to noise” ratio (SNR).

How do we know that our results are not due to chance?

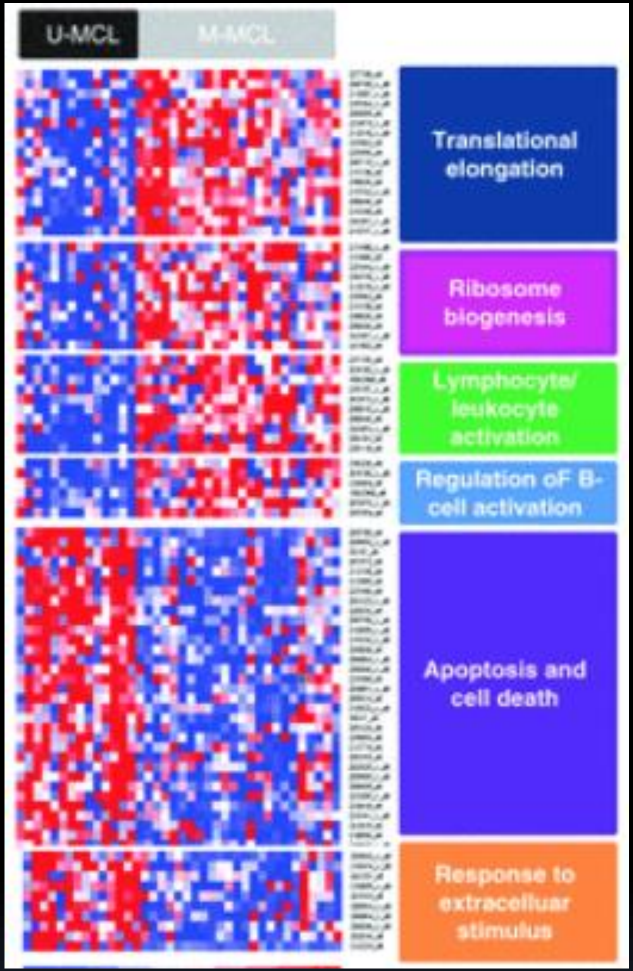
- The resulting distribution of the randomized SNR (which we call the Null distribution) will be something like this



- The blue area is the p-value, and tells me the probability of observing a SNR as the one I had in my experiment, if the labels were random. If the p-value is very small, I reject the null hypothesis.

Algorithms Diff Exp for RNA-Seq

- There are a new suite of algorithms that find differential expression in RNA seq. They use different statistical assumptions that are specific to the digital nature of the data.
  - Cufflinks
  - DESeq from Wolfgang Huber
  - EdgeR from Gordon Smyth
  - Limma
- Many use a type of statistical model called Generalized Linear Models (GLM)
- These still need systematic evaluation

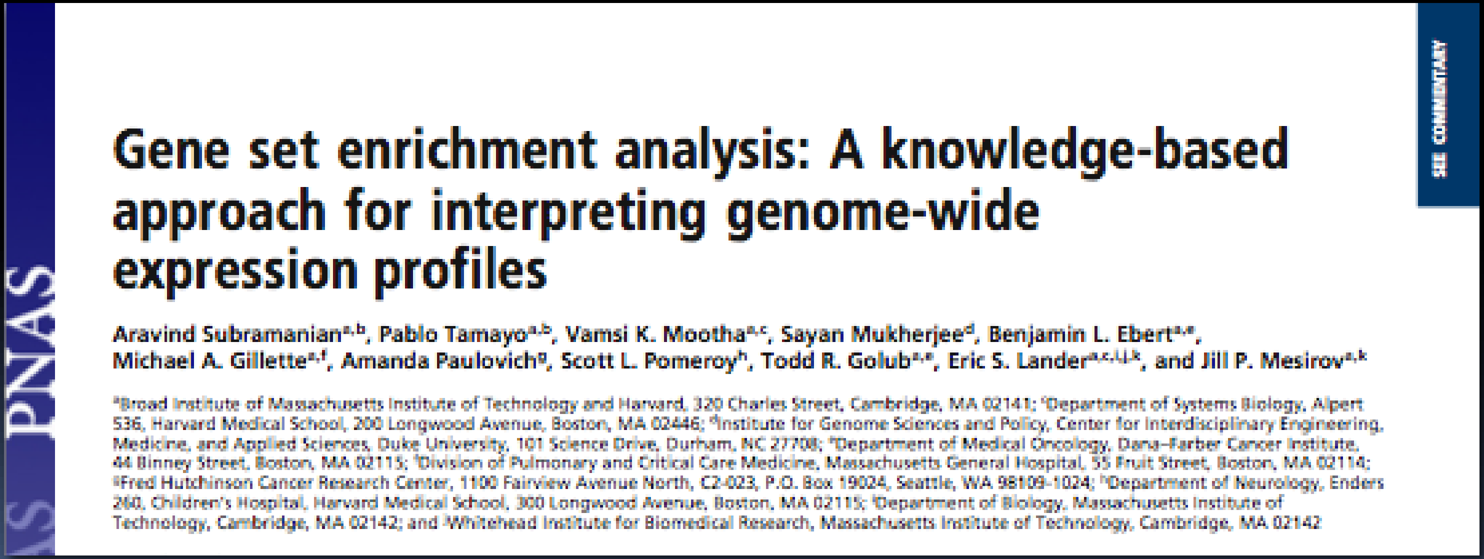


Assessing Biological significance

Interpreting the results of differential expression

- How can we assign a biological interpretation to the list of genes that we obtained using differential expression?

- A god idea came in 2005 with this paper (8600 citations so far)



PNAS
SEE COMMENTARY

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian[a,b], Pablo Tamayo[a,b], Vamsi K. Mootha[a,c], Sayan Mukherjee[d], Benjamin L. Ebert[a,e], Michael A. Gillette[a,f], Amanda Paulovich[g], Scott L. Pomeroy[h], Todd R. Golub[a,e], Eric S. Lander[a,c,i,j,k], and Jill P. Mesirov[a,k]

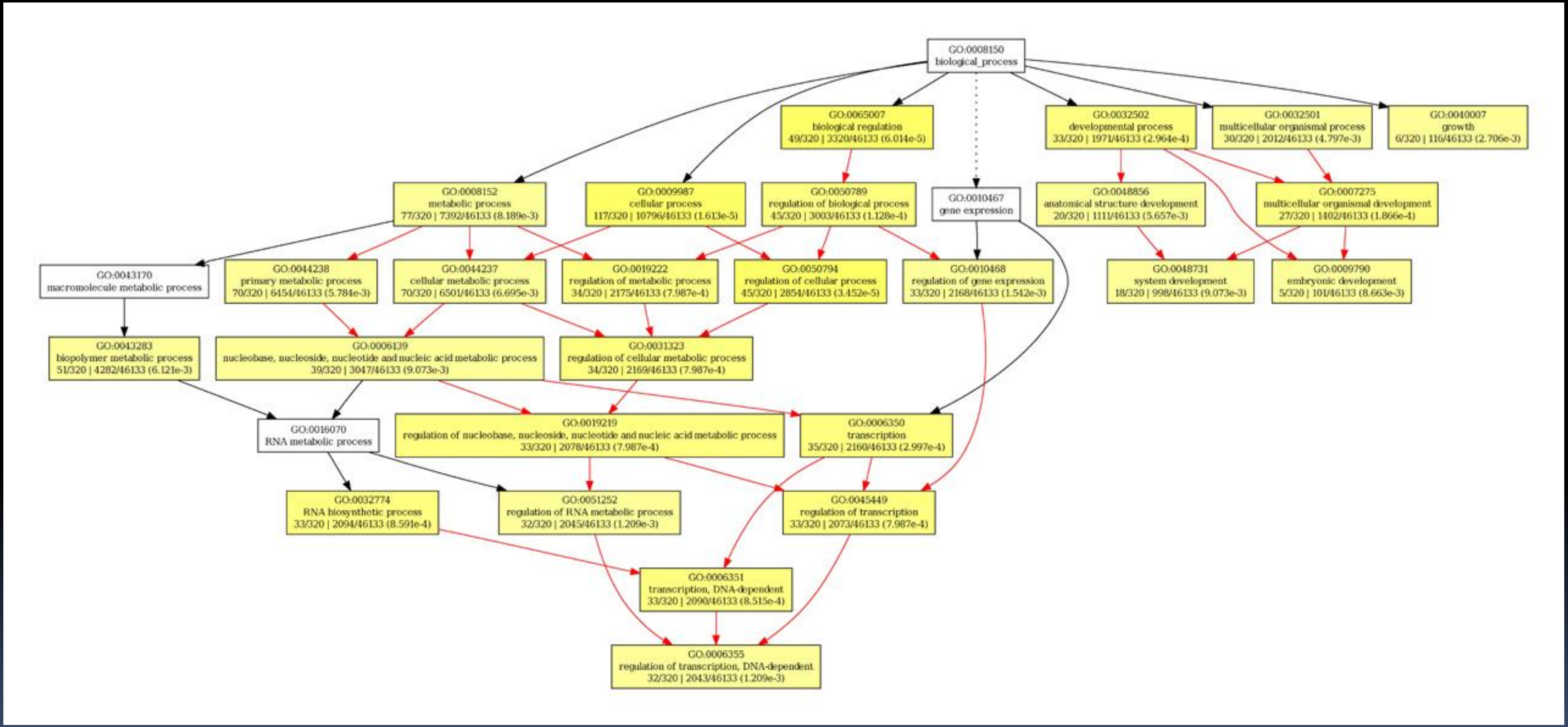
[a]Broad Institute of Massachusetts Institute of Technology and Harvard, 320 Charles Street, Cambridge, MA 02141; [c]Department of Systems Biology, Alpert 536, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02446; [d]Institute for Genome Sciences and Policy, Center for Interdisciplinary Engineering, Medicine, and Applied Sciences, Duke University, 101 Science Drive, Durham, NC 27708; [e]Department of Medical Oncology, Dana–Farber Cancer Institute, 44 Binney Street, Boston, MA 02115; [f]Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114; [g]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, C2-023, P.O. Box 19024, Seattle, WA 98109-1024; [h]Department of Neurology, Enders 260, Children's Hospital, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115; [i]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142; and [j]Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, MA 02142

Interpreting the results of differential expression

- The algorithm proposed is called GSEA, and uses prior knowledge (Gene Sets) contrasted with the list of differentially expressed genes

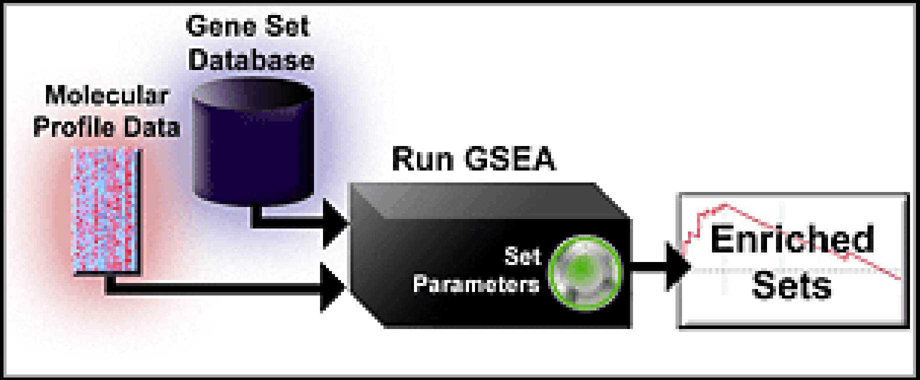
- Which prior knowledge?
  - MSigDB (Molecular Signatures DB) ~13000 gene signatures
    - http://software.broadinstitute.org/gsea/msigdb/index.jsp
  - BioCARTA pathways: http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
  - Gene Ontology

A chunk of Gene Ontology


GO:0008150 biological_process
GO:0065007 biological regulation 49/320 | 3320/46133 (6.014e-5)
GO:0032502 developmental process 33/320 | 1971/46133 (2.964e-4)
GO:0032501 multicellular organismal process 30/320 | 2012/46133 (4.797e-3)
GO:0040007 growth 6/320 | 116/46133 (2.706e-3)
GO:0008152 metabolic process 77/320 | 7392/46133 (8.189e-3)
GO:0009987 cellular process 117/320 | 10796/46133 (1.613e-5)
GO:0050789 regulation of biological process 45/320 | 3003/46133 (1.128e-4)
GO:0010467 gene expression
GO:0048856 anatomical structure development 20/320 | 1111/46133 (5.657e-3)
GO:0007275 multicellular organismal development 27/320 | 1402/46133 (1.866e-4)
GO:0043170 macromolecule metabolic process
GO:0044238 primary metabolic process 70/320 | 6454/46133 (5.784e-3)
GO:0044237 cellular metabolic process 70/320 | 6501/46133 (6.695e-3)
GO:0019222 regulation of metabolic process 34/320 | 2175/46133 (7.987e-4)
GO:0050794 regulation of cellular process 45/320 | 2854/46133 (3.452e-5)
GO:0010468 regulation of gene expression 33/320 | 2168/46133 (1.542e-3)
GO:0048731 system development 18/320 | 998/46133 (9.073e-3)
GO:0009790 embryonic development 5/320 | 101/46133 (8.663e-3)
GO:0043283 biopolymer metabolic process 51/320 | 4282/46133 (6.121e-3)
GO:0006139 nucleobase, nucleoside, nucleotide and nucleic acid metabolic process 39/320 | 3047/46133 (9.073e-3)
GO:0031323 regulation of cellular metabolic process 34/320 | 2169/46133 (7.987e-4)
GO:0016070 RNA metabolic process
GO:0019219 regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process 33/320 | 2078/46133 (7.987e-4)
GO:0006350 transcription 35/320 | 2160/46133 (2.997e-4)
GO:0032774 RNA biosynthetic process 33/320 | 2094/46133 (8.591e-4)
GO:0051252 regulation of RNA metabolic process 32/320 | 2045/46133 (1.209e-3)
GO:0045449 regulation of transcription 33/320 | 2073/46133 (7.987e-4)
GO:0006351 transcription, DNA-dependent 33/320 | 2090/46133 (8.515e-4)
GO:0006355 regulation of transcription, DNA-dependent 32/320 | 2043/46133 (1.209e-3)

Interpreting the results of differential expression

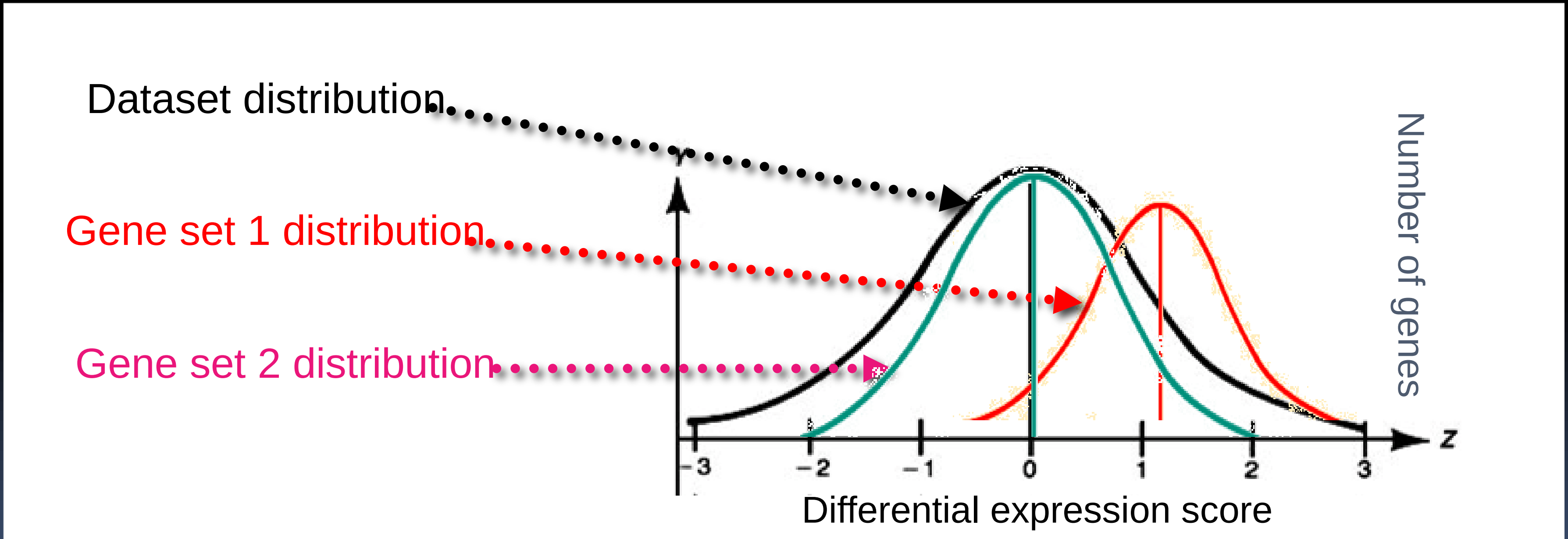
GSEA applies Kolmogorov-Smirnof test to find assymmetrical distributions for defined blocks of genes in datasets whole distribution.



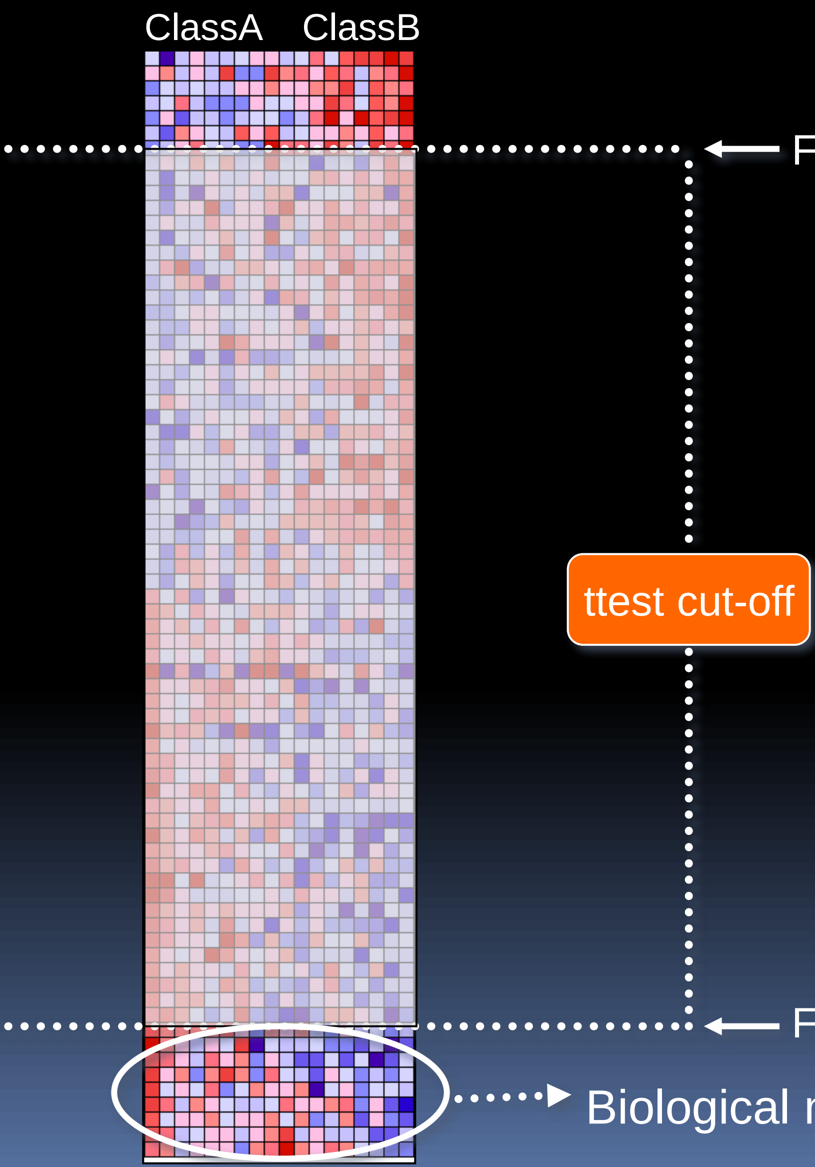
Is this particular Gene Set enriched in my experiment?

Interpreting the results of differential expression

The Kolmogorov–Smirnov test is used to determine whether two probability distributions differ.



Interpreting the results of differential expression

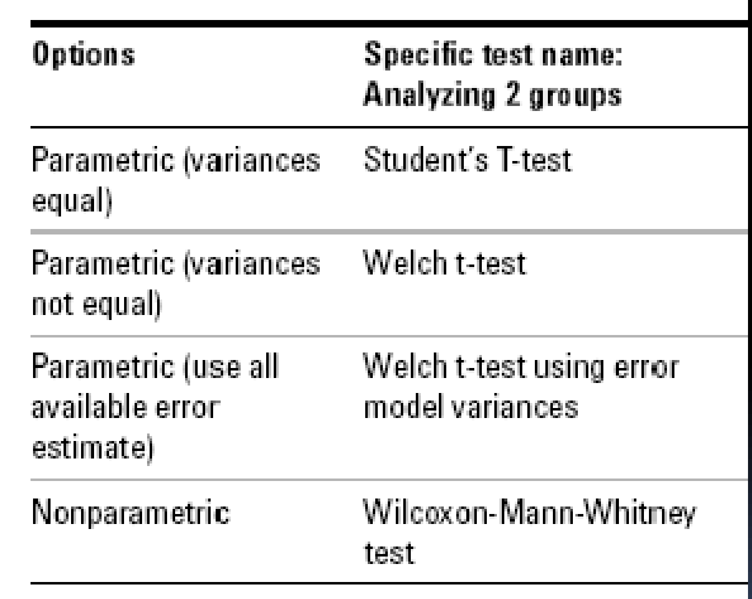


FDR<0.05

FDR<0.05

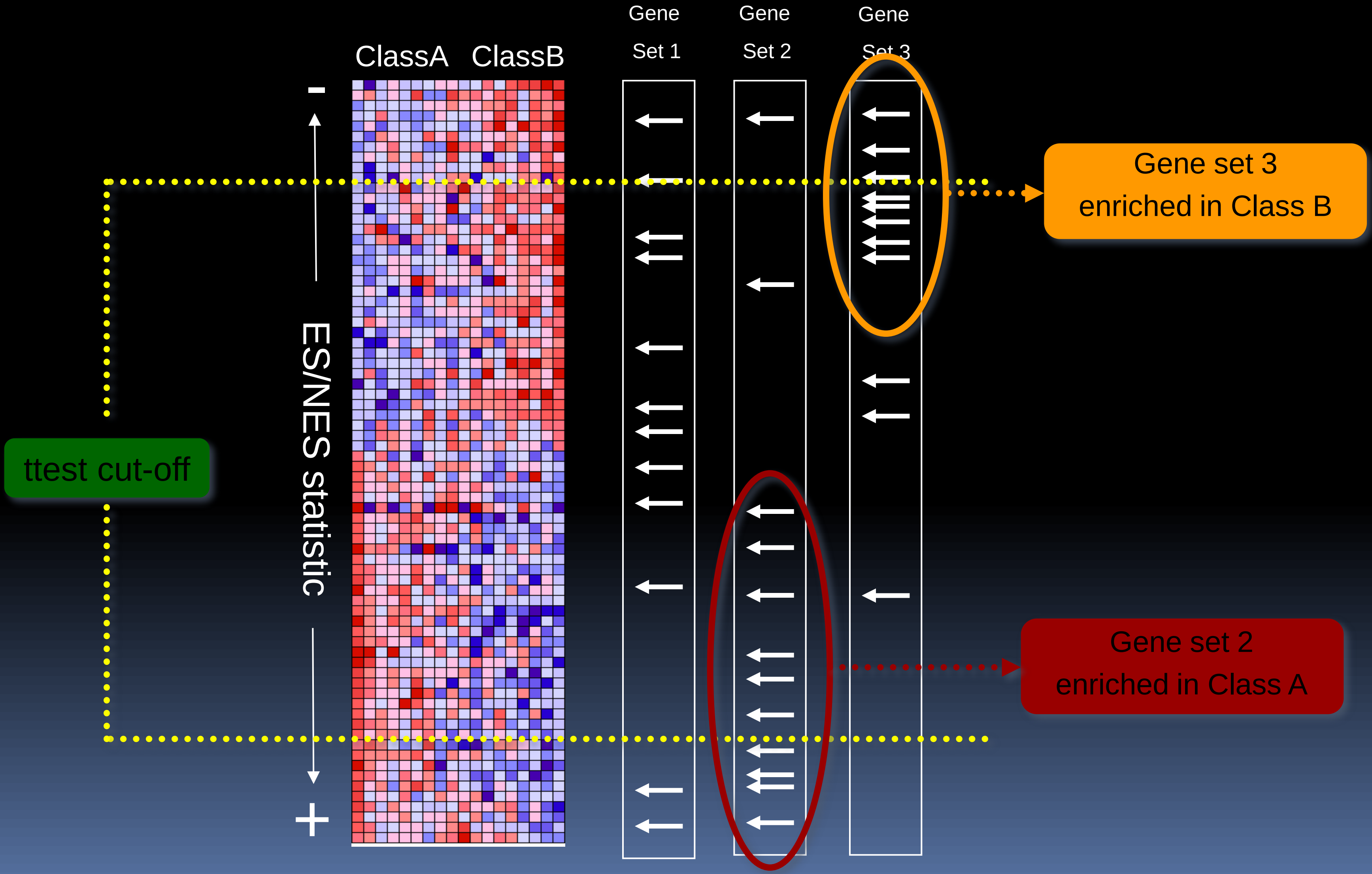
Biological meaning?

...testing genes independently...



Options	Specific test name: Analyzing 2 groups
Parametric (variances equal)	Student's T-test
Parametric (variances not equal)	Welch t-test
Parametric (use all available error estimate)	Welch t-test using error model variances
Nonparametric	Wilcoxon-Mann-Whitney test

Interpreting the results of differential expression



GSEA: Key Features

- We rank all genes based on their differential expression score

- We identifies gene sets whose member genes are clustered either towards top or bottom of the ranked list (i.e. up- or down regulated)

- We compute an enrichment score for each gene set

- We do a permutation test to identify significantly enriched categories

GSEA Algorithm: Definition of Enrichment Scores

The equations

w_j = measure of differential expression of gene *j* between group A and group B

1. **Order the genes in a ranked list L so w_j decreases from the top (j=1) to the bottom (j=N) of the list**

2. **Account for the locations of the genes in Gene Set *S* (“hits”) weighted by w_j and the locations of genes not in *S* (“misses”) from the top of the list down to a given position *i* in *L***

$$K_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|w_j|^t}{\Sigma}$$

where

for GSEA the default is *t* = 1, for Kolmogorov-Smirnov *t* = 0

$$K_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)}$$

N_H = # genes in S

N = # genes in platform

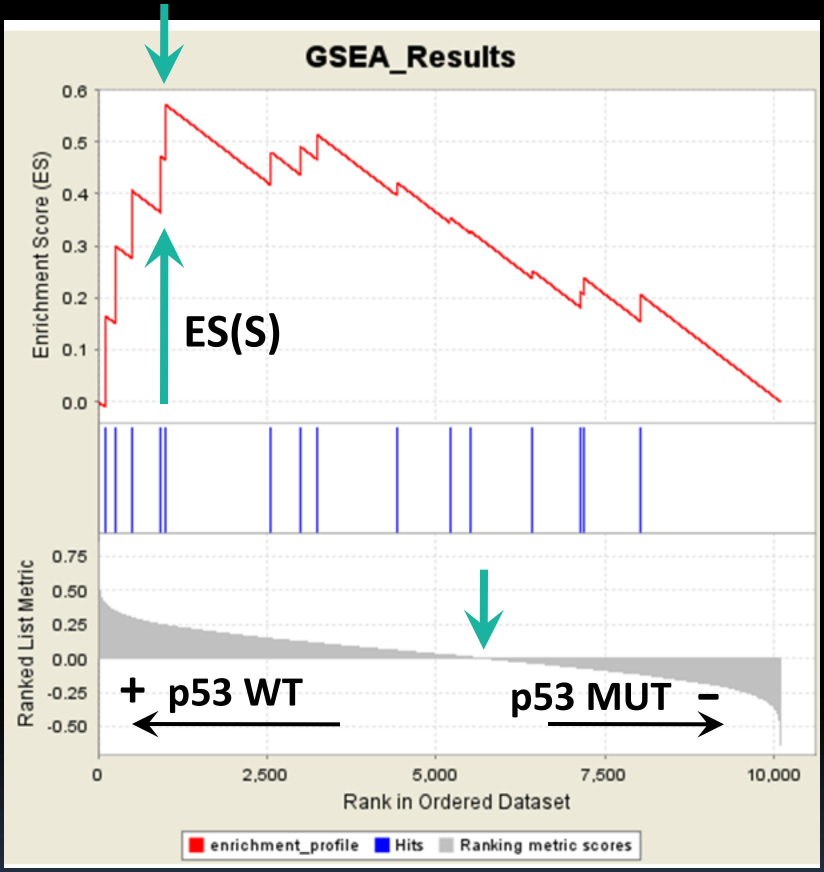
3. **Calculate maximum deviation from zero of K_{hit} - K_{miss}** over $1 \leq i \leq N$:

$ES(S,i) = K_{hit}(S,i) - K_{miss}(S,i)$ Note $K_{hit}(S,N) = K_{miss}(S,N) = 1$ so ES(S,N) = 0

$ES(S)$ = max deviation{$ES(S,i)$} (greatest excursion of the ES(S,i) from 0)

The running enrichment score for a positive ES gene set from the P53 GSEA example data set

running enrichment score

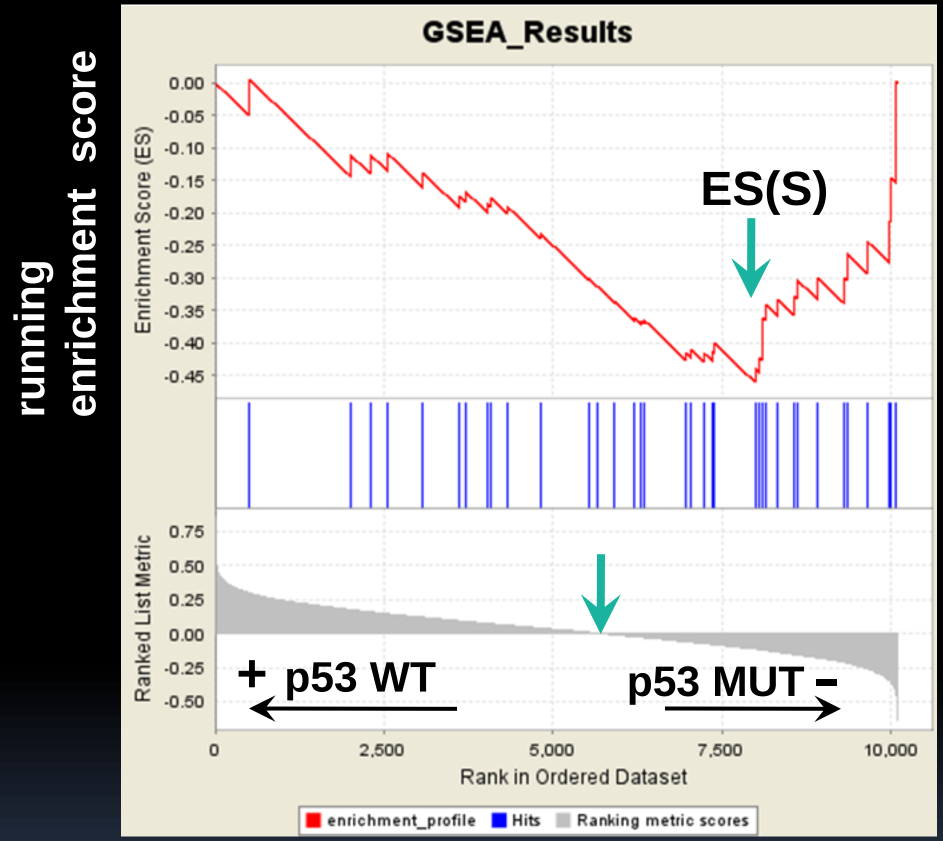


locations of genes in S

Zero crossing of ranking metric values

underlying running enrichment score figure copied from
http://www.broadinstitute.org/gsea/datasets.jsp
p53 dataset (gene set is lairPathway)

The running enrichment score for a negative ES gene set from the P53 GSEA example data set



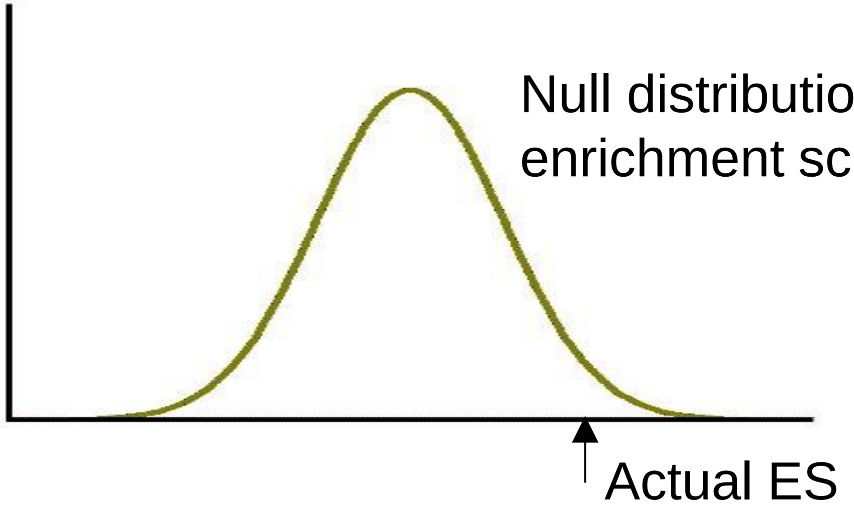
locations of genes in S

Zero crossing of ranking metric values

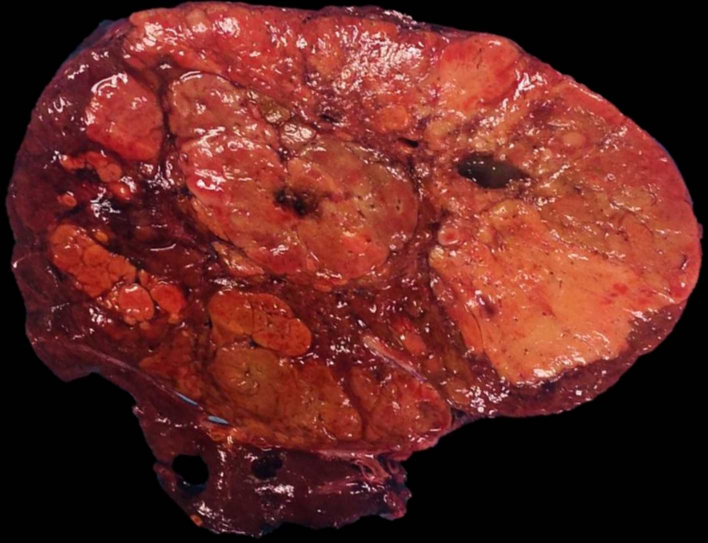
running enrichment score figure copied from
http://www.broadinstitute.org/gsea/datasets.jsp
p53 dataset (gene set is BRCA_UP)

GSEA: Permutation Test

- Randomize data (groups), rank genes again and repeat test 1000 times
- Null distribution of 1000 ES for geneset



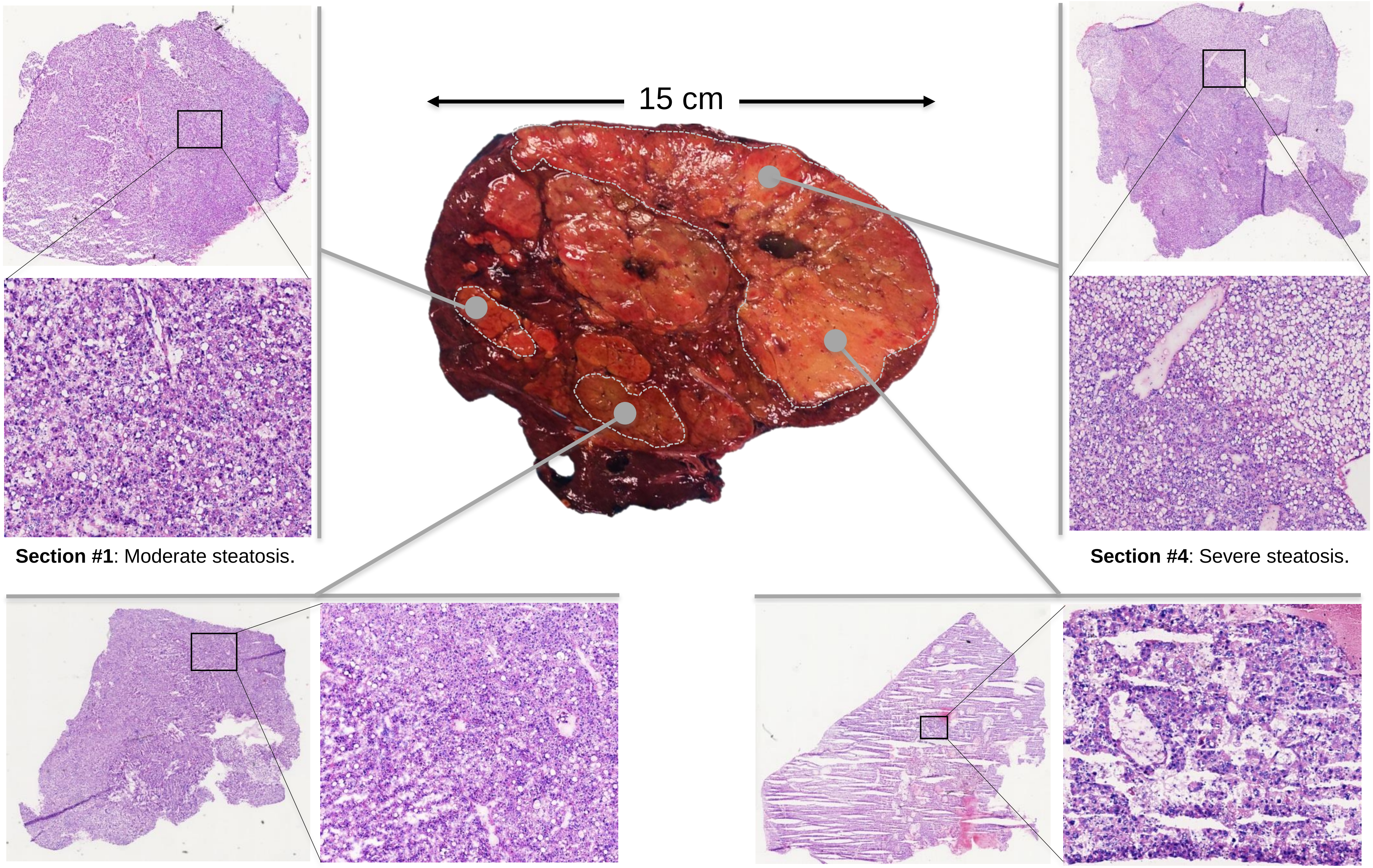
- FDR q-value computed – corrected for gene set size and testing multiple gene sets



Characterization of Intra-Tumor Heterogeneity in Hepatocellular Carcinoma

Intra-tumoral Heterogeneity

Histology



Section #1: Moderate steatosis.

Section #4: Severe steatosis.

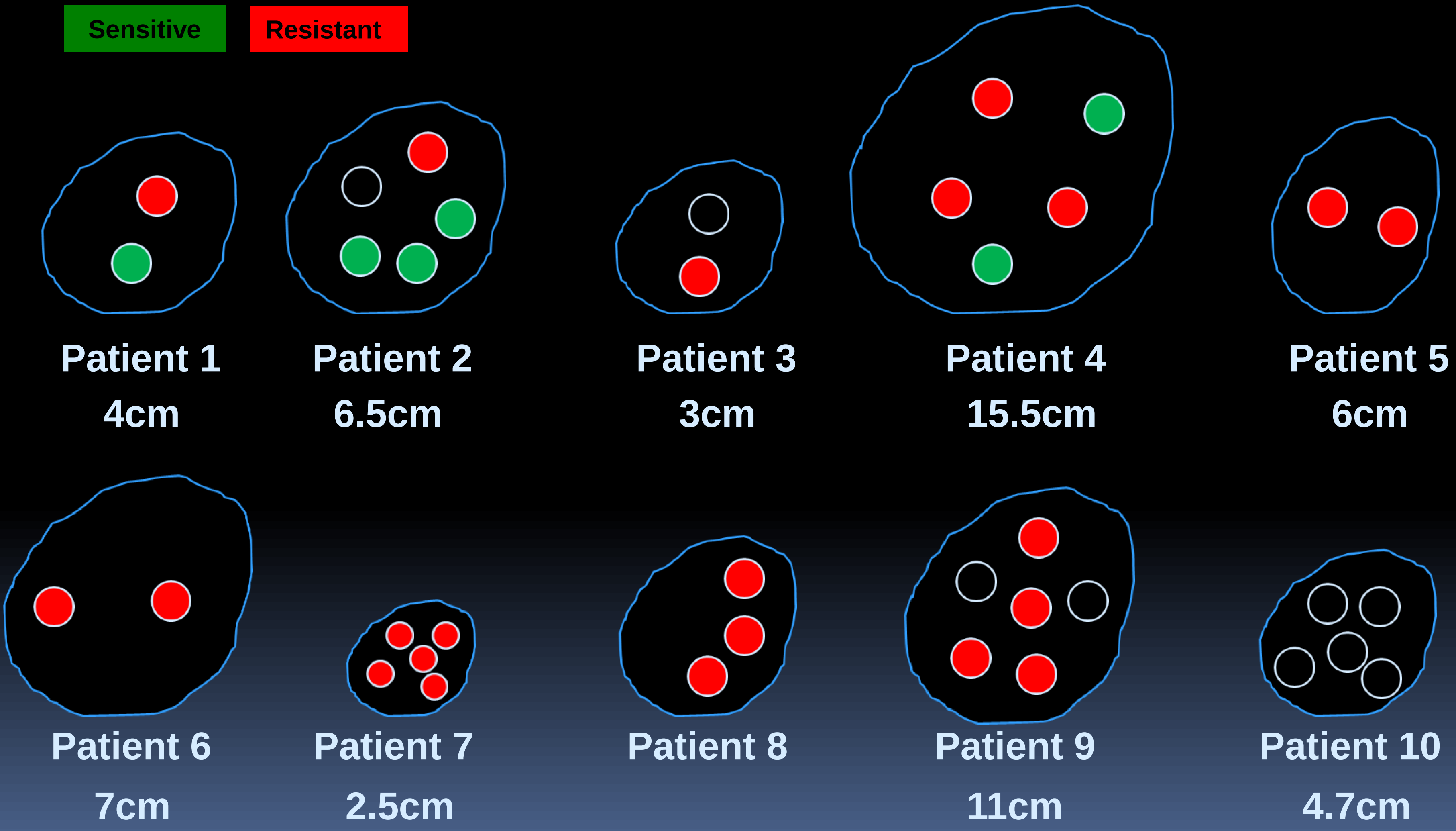
Section #2: Moderate steatosis.

Section #3: No steatosis. Bile production

Nearest Template Prediction

Sorafenib Resistance Signature

🟩 Sensitive 🟥 Resistant



Patient 1
4cm

Patient 2
6.5cm

Patient 3
3cm

Patient 4
15.5cm

Patient 5
6cm

Patient 6
7cm

Patient 7
2.5cm

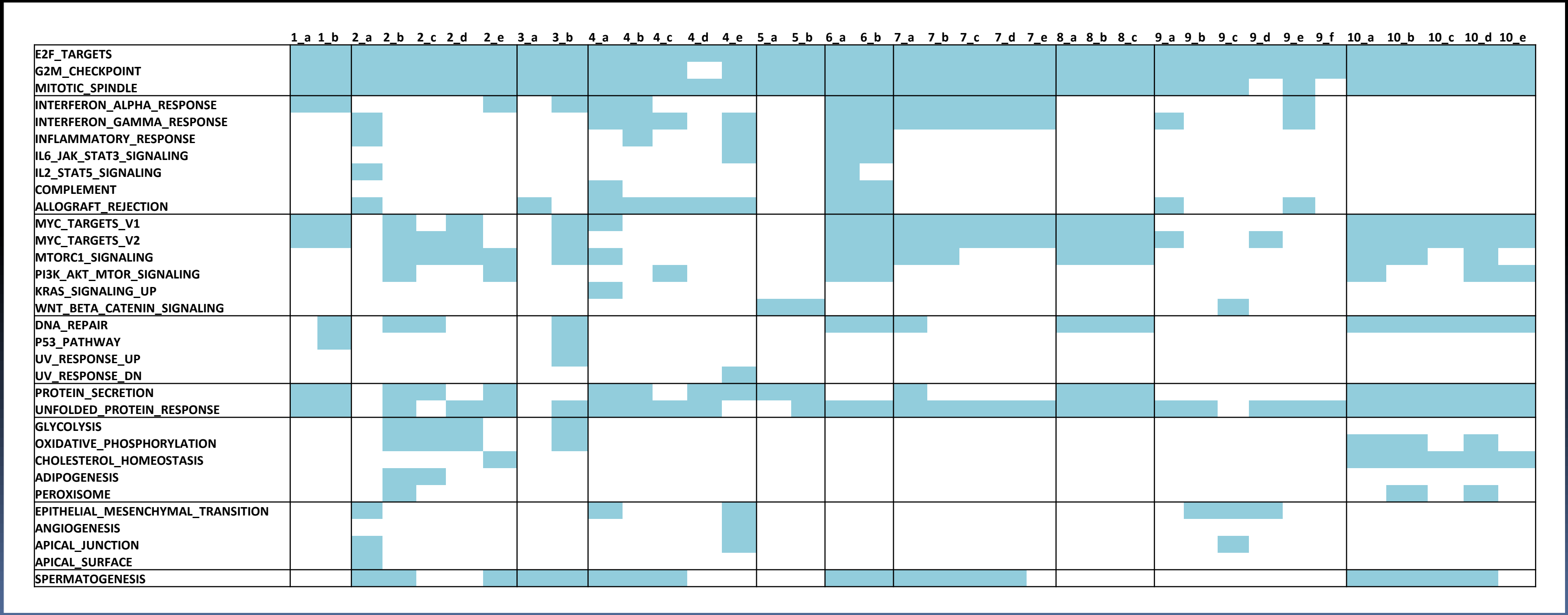
Patient 8

Patient 9
11cm

Patient 10
4.7cm

Gene Set Enrichment Analysis By Patient

- Proliferation up in every sample
- Immune, MTOR signaling, and metabolism, migration processes are variable
- MYC signaling, DNA repair, protein pathways, and spermatogenesis are more homogeneous

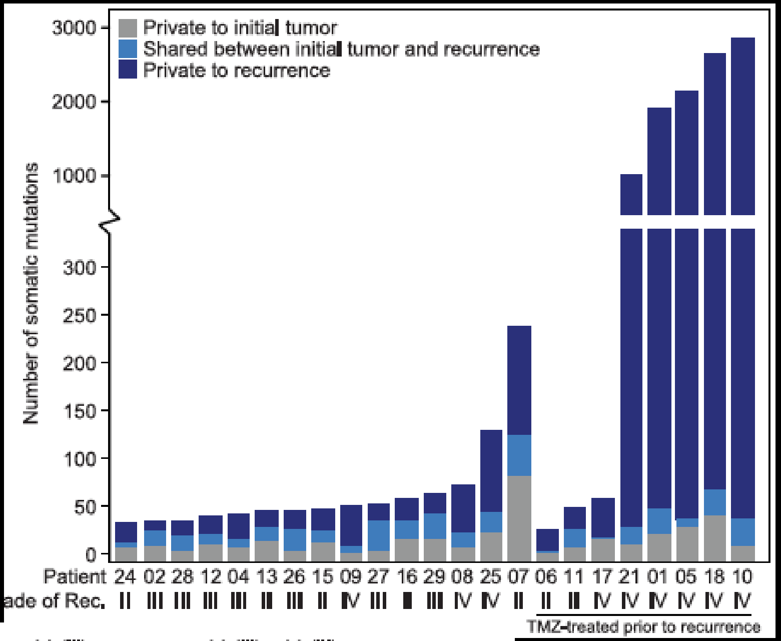


	1_a	1_b	2_a	2_b	2_c	2_d	2_e	3_a	3_b	4_a	4_b	4_c	4_d	4_e	5_a	5_b	6_a	6_b	7_a	7_b	7_c	7_d	7_e	8_a	8_b	8_c	9_a	9_b	9_c	9_d	9_e	9_f	10_a	10_b	10_c	10_d	10_e
E2F_TARGETS	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
G2M_CHECKPOINT	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
MITOTIC_SPINDLE	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	■	■	■	■
INTERFERON_ALPHA_RESPONSE	■	■					■		■	■	■						■	■	■	■	■	■	■								■						
INTERFERON_GAMMA_RESPONSE			■							■	■	■		■			■	■	■	■	■	■	■				■				■						
INFLAMMATORY_RESPONSE			■								■			■			■	■																			
IL6_JAK_STAT3_SIGNALING														■			■	■																			
IL2_STAT5_SIGNALING			■														■																				
COMPLEMENT										■							■	■																			
ALLOGRAFT_REJECTION			■					■		■	■	■	■	■			■	■									■				■						
MYC_TARGETS_V1	■	■		■		■			■	■							■	■	■	■	■	■	■	■	■	■							■	■	■	■	■
MYC_TARGETS_V2	■	■		■	■	■			■								■	■	■	■	■	■	■	■	■	■	■			■			■	■	■	■	■
MTORC1_SIGNALING				■	■	■	■		■	■							■	■	■	■				■	■	■							■	■		■	
PI3K_AKT_MTOR_SIGNALING				■			■					■					■	■															■			■	■
KRAS_SIGNALING_UP										■																											
WNT_BETA_CATENIN_SIGNALING															■	■													■								
DNA_REPAIR		■		■	■				■								■	■	■					■	■	■							■	■	■	■	■
P53_PATHWAY		■							■																												
UV_RESPONSE_UP									■																												
UV_RESPONSE_DN														■																							
PROTEIN_SECRETION	■	■		■	■		■			■	■		■	■	■	■			■					■	■	■							■	■	■	■	■
UNFOLDED_PROTEIN_RESPONSE	■	■		■		■	■		■	■	■	■	■			■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■
GLYCOLYSIS				■	■	■			■																												
OXIDATIVE_PHOSPHORYLATION				■	■	■			■																								■	■		■	
CHOLESTEROL_HOMEOSTASIS							■																										■	■	■	■	■
ADIPOGENESIS				■	■																																
PEROXISOME				■																														■		■	
EPITHELIAL_MESENCHYMAL_TRANSITION			■							■				■														■	■	■							
ANGIOGENESIS														■																							
APICAL_JUNCTION			■											■															■								
APICAL_SURFACE			■																																		
SPERMATOGENESIS			■	■			■	■	■	■	■	■					■	■	■	■	■	■											■	■	■	■	

More on Cancer Heterogeneity

www.sciencemag.org SCIENCE VOL 343 10 JANUARY 2014

Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma



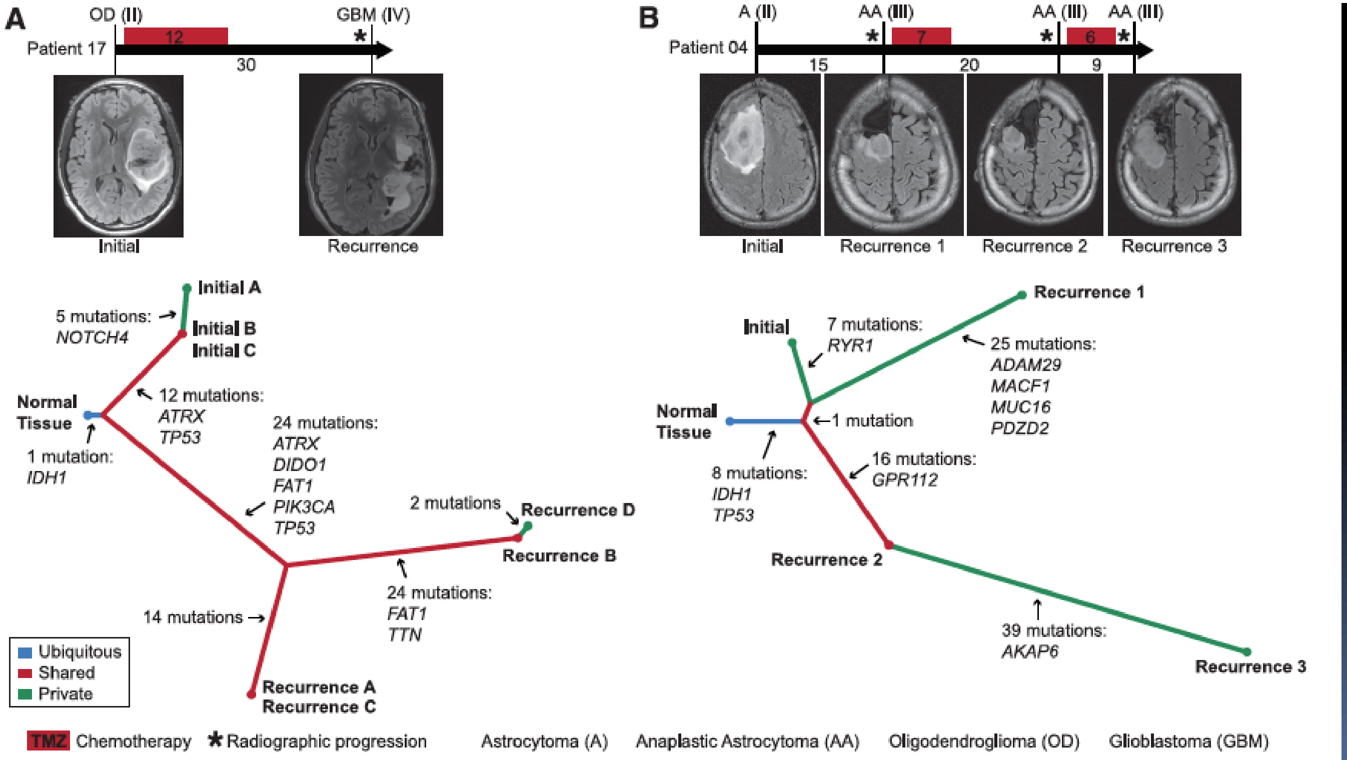


Fig. 2. Temporal and spatial patterns of clonal evolution in the tumors of two glioma patients. (**A** and **B**) A timeline of treatment histories for patient 17 (A) and patient 04 (B) (top, intervals labeled in months). Vertical bars correspond to the time of tumor resection and are labeled with the tumor diagnosis and grade. Representative MRIs are also shown. A phylogenetic tree (bottom) depicts the patterns of clonal evolution of these tumors inferred from the pattern and frequency of somatic mutations, highlighting genes frequently mutated in cancer.

Proposed homework

Read: Subramanian A, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50

Or

Read: Burrell RA1, McGranahan N, Bartek J, Swanton C., The causes and consequences of genetic heterogeneity in cancer evolution, Nature. 2013 Sep 19;501(7467):338-45

Or

Explore MsigDB (http://software.broadinstitute.org/gsea/msigdb/index.jsp)

at the Broad Institute