Gene Set Enrichment Analysis

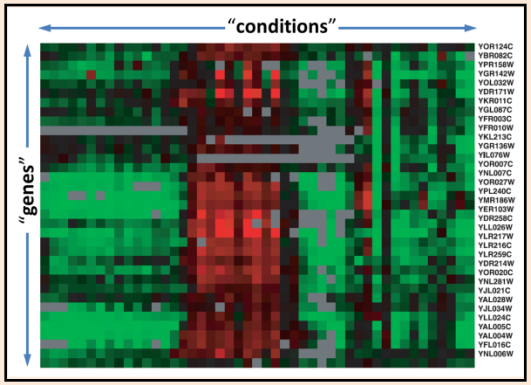
Genome 373

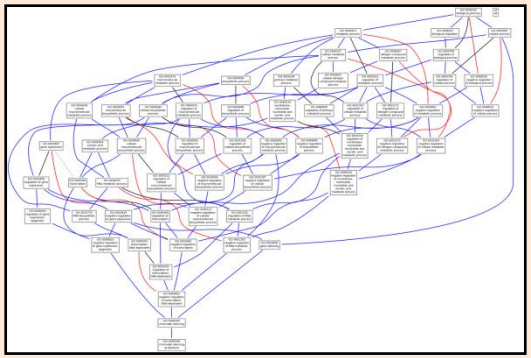
Genomic Informatics

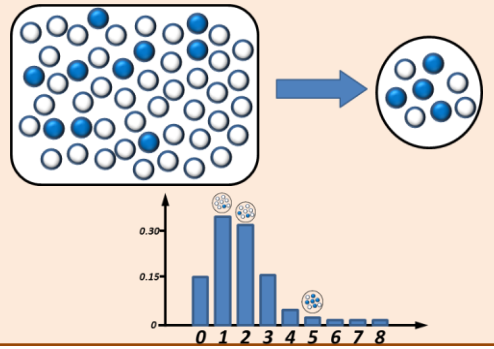
Elhanan Borenstein

A quick review

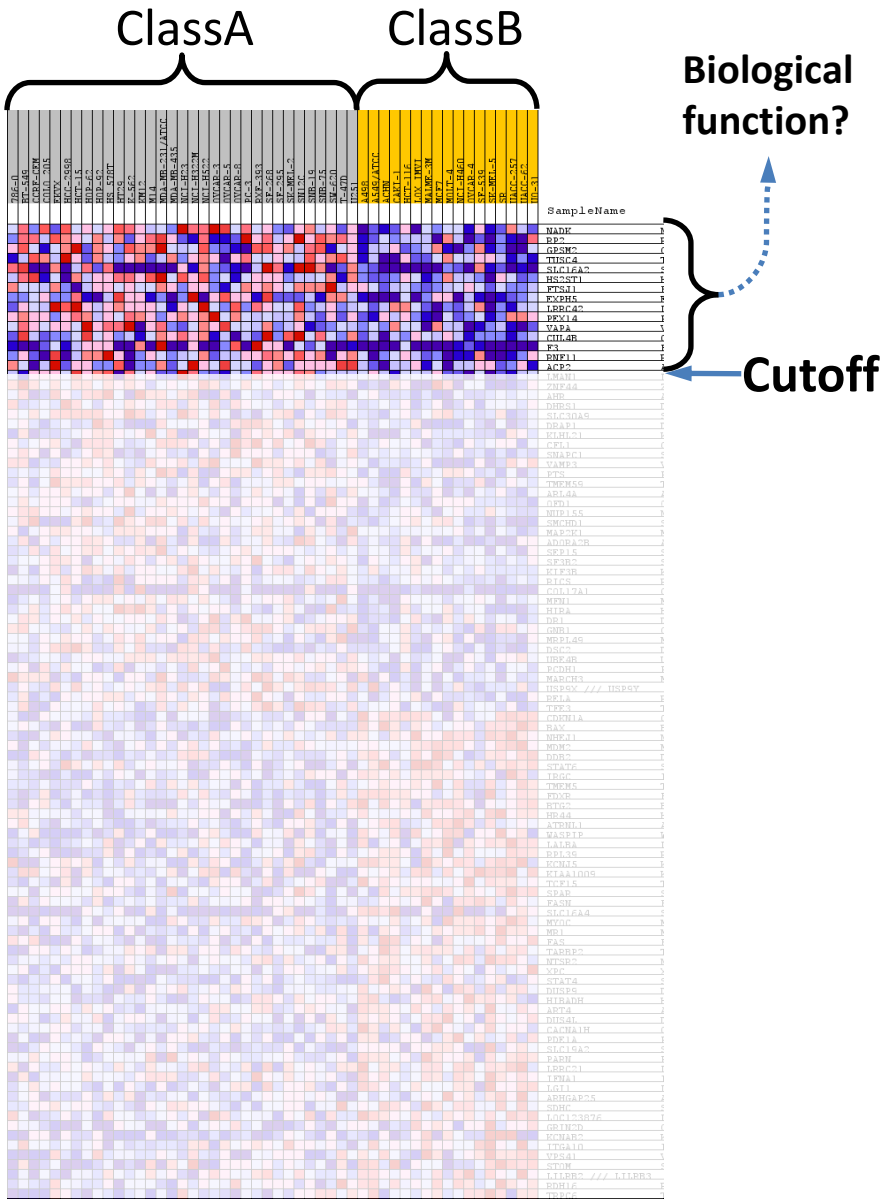
- Gene expression profiling
  - Which molecular processes/functions are involved in a certain phenotype (e.g., disease, stress response, etc.)
- The Gene Ontology (GO) Project
  - Provides shared vocabulary/annotation
  - Terms are linked in a complex structure
- Enrichment analysis:
  - Find the “most” differentially expressed genes
  - Identify **over-represented** annotations
  - Modified Fisher's exact test

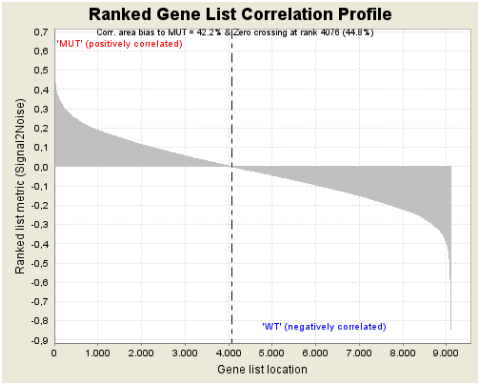




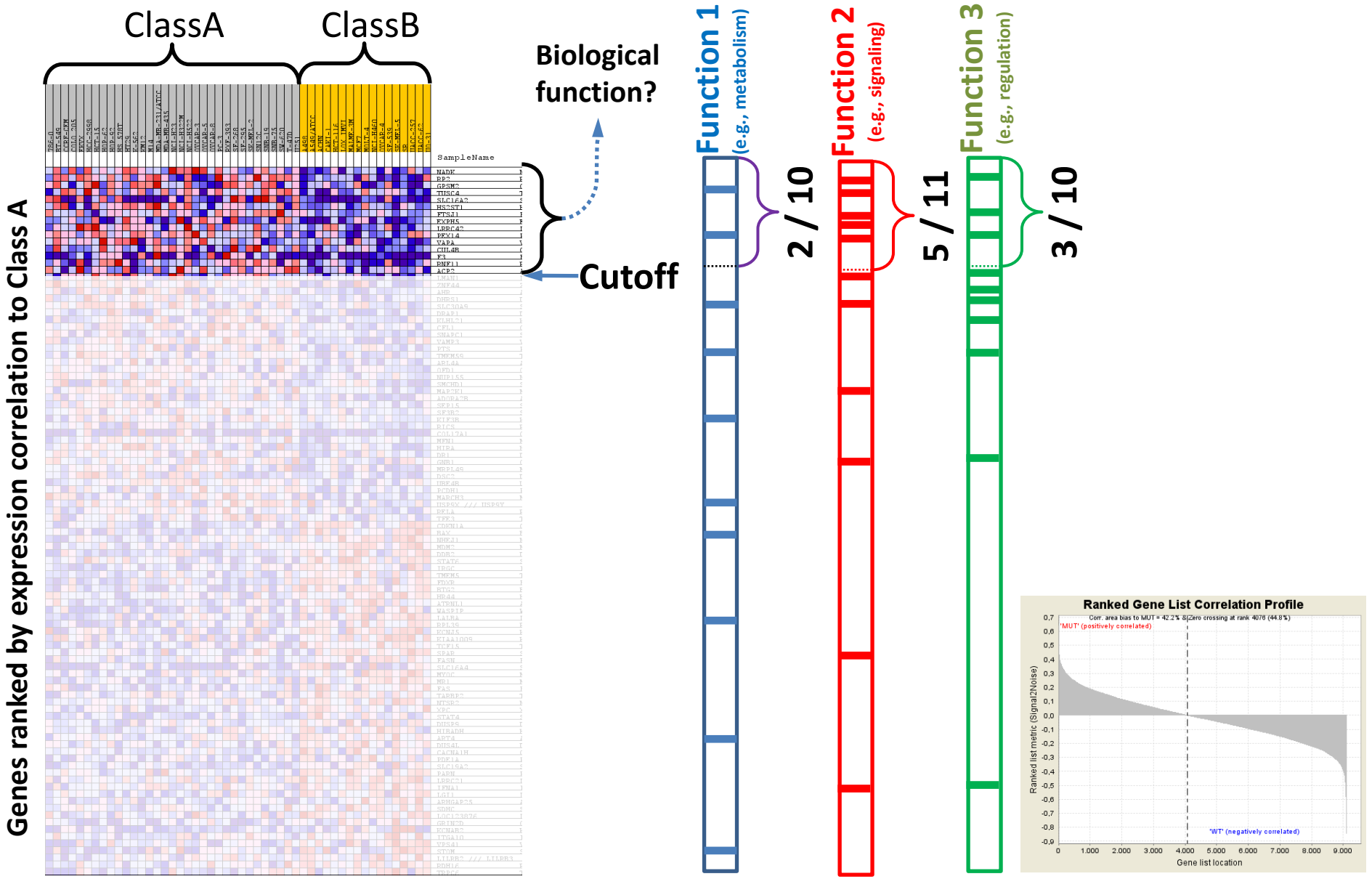


Enrichment Analysis



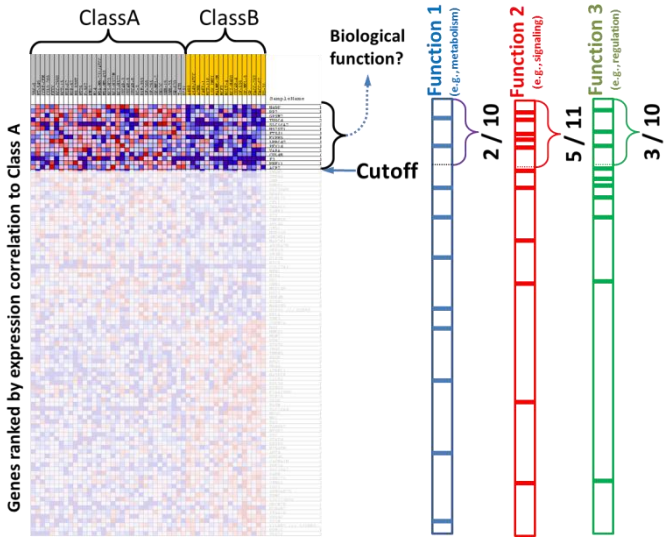


Enrichment Analysis



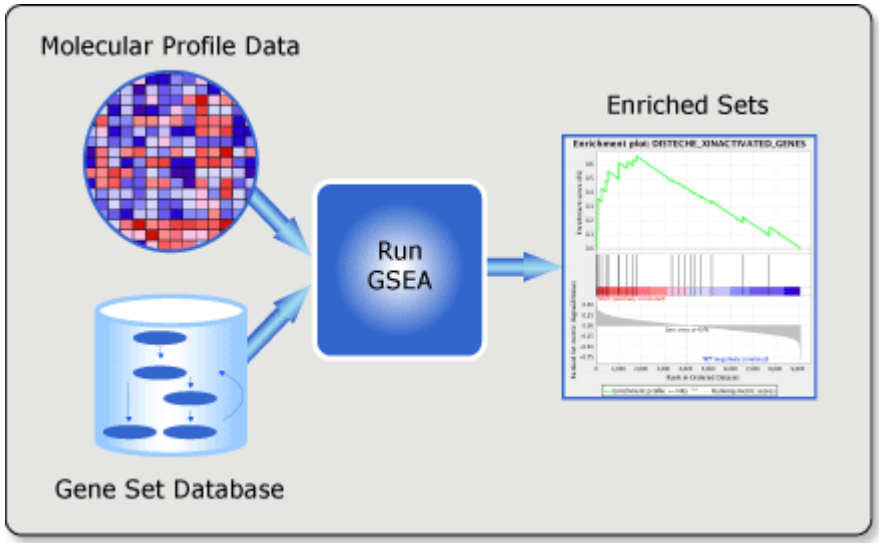
Problems with cutoff-based analysis

- After correcting for multiple hypotheses testing, no individual gene may meet the threshold due to noise.
- Alternatively, one may be left with a long list of significant genes without any unifying biological theme.
- The cutoff value is often arbitrary!
- **We are really examining only a handful of genes, totally ignoring much of the data**



Gene Set Enrichment Analysis

- MIT, Broad Institute
- V 2.0 available since Jan 2007



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

Contributed by Eric S. Lander, August 2, 2005

Although genomewide RNA expression analysis has become a routine tool in biomedical research, extracting biological insight from such information remains a major challenge. Here, we describe a powerful analytical method called Gene Set Enrichment Analysis (GSEA) for interpreting gene expression data. The method derives its power by focusing on gene sets, that is, groups of genes that share common biological function, chromosomal location, or regulation. We demonstrate how GSEA yields insights into several cancer-related data sets, including leukemia and lung cancer. Notably, where single-gene analysis finds little similarity between two independent studies of patient survival in lung cancer, GSEA reveals many biological pathways in common. The GSEA method is embodied in a freely available software package, together with an initial database of 1,325 biologically defined gene sets.

microarray

Genomewide expression analysis with DNA microarrays has become a mainstay of genomics research (1, 2). The challenge no longer lies in obtaining gene expression profiles, but rather in interpreting the results to gain insights into biological mechanisms.

In a typical experiment, mRNA expression profiles are generated for thousands of genes from a collection of samples belonging to one of two classes, for example, tumors that are sensitive vs. resistant to a drug. The genes can be ordered in a ranked list *L*, according to their differential expression between the classes. The challenge is to extract meaning from this list.

A common approach involves focusing on a handful of genes at the top and bottom of *L* (i.e., those showing the largest difference) to discern telltale biological clues. This approach has a few major limitations.

(*i*) After correcting for multiple hypotheses testing, no individual gene may meet the threshold for statistical significance, because the relevant biological differences are modest relative to the noise inherent to the microarray technology.

(*ii*) Alternatively, one may be left with a long list of statistically significant genes without any unifying biological theme. Interpretation can be daunting and ad hoc, being dependent on a biologist's area of expertise.

(*iii*) Single-gene analysis may miss important effects on pathways. Cellular processes often affect sets of genes acting in concert. An increase of 20% in all genes encoding members of a metabolic pathway may dramatically alter the flux through the pathway and may be more important than a 20-fold increase in a single gene.

(*iv*) When different groups study the same biological system, the list of statistically significant genes from the two studies may show distressingly little overlap (3).

To overcome these analytical challenges, we recently developed a method called Gene Set Enrichment Analysis (GSEA) that evaluates microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or coexpression in previous experiments. The goal of GSEA is to determine whether members of a gene set *S* tend to occur toward the top (or bottom) of the list *L*, in which case the gene set is correlated with the phenotypic class distinction.

We used a preliminary version of GSEA to analyze data from muscle biopsies from diabetics vs. healthy controls (4). The method revealed that genes involved in oxidative phosphorylation show reduced expression in diabetics, although the average decrease per gene is only 20%. The results from this study have been independently validated by other microarray studies (5) and by *in vivo* functional studies (6).

Given this success, we have developed GSEA into a robust technique for analyzing molecular profiling data. We studied its characteristics and performance and substantially revised and generalized the original method for broader applicability.

In this paper, we provide a full mathematical description of the GSEA methodology and illustrate its utility by applying it to several diverse biological problems. We have also created a software package, called GSEA-P and an initial inventory of gene sets (Molecular Signature Database, MSigDB), both of which are freely available.

Methods

Overview of GSEA. GSEA considers experiments with genomewide expression profiles from samples belonging to two classes, labeled 1 or 2. Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric (Fig. 1*A*).

Given an *a priori* defined set of genes *S* (e.g., genes encoding products in a metabolic pathway, located in the same cytogenetic band, or sharing the same GO category), the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout *L* or primarily found at the top or bottom. We expect

Freely available online through the PNAS open access option.

Abbreviations: ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; *ES*, enrichment score; FDR, false discovery rate; GSEA, Gene Set Enrichment Analysis; MAPK, mitogen-activated protein kinase; MSigDB, Molecular Signature Database; *NES*, normalized enrichment score.

See Commentary on page 15278.

[†]A.S. and P.T. contributed equally to this work.

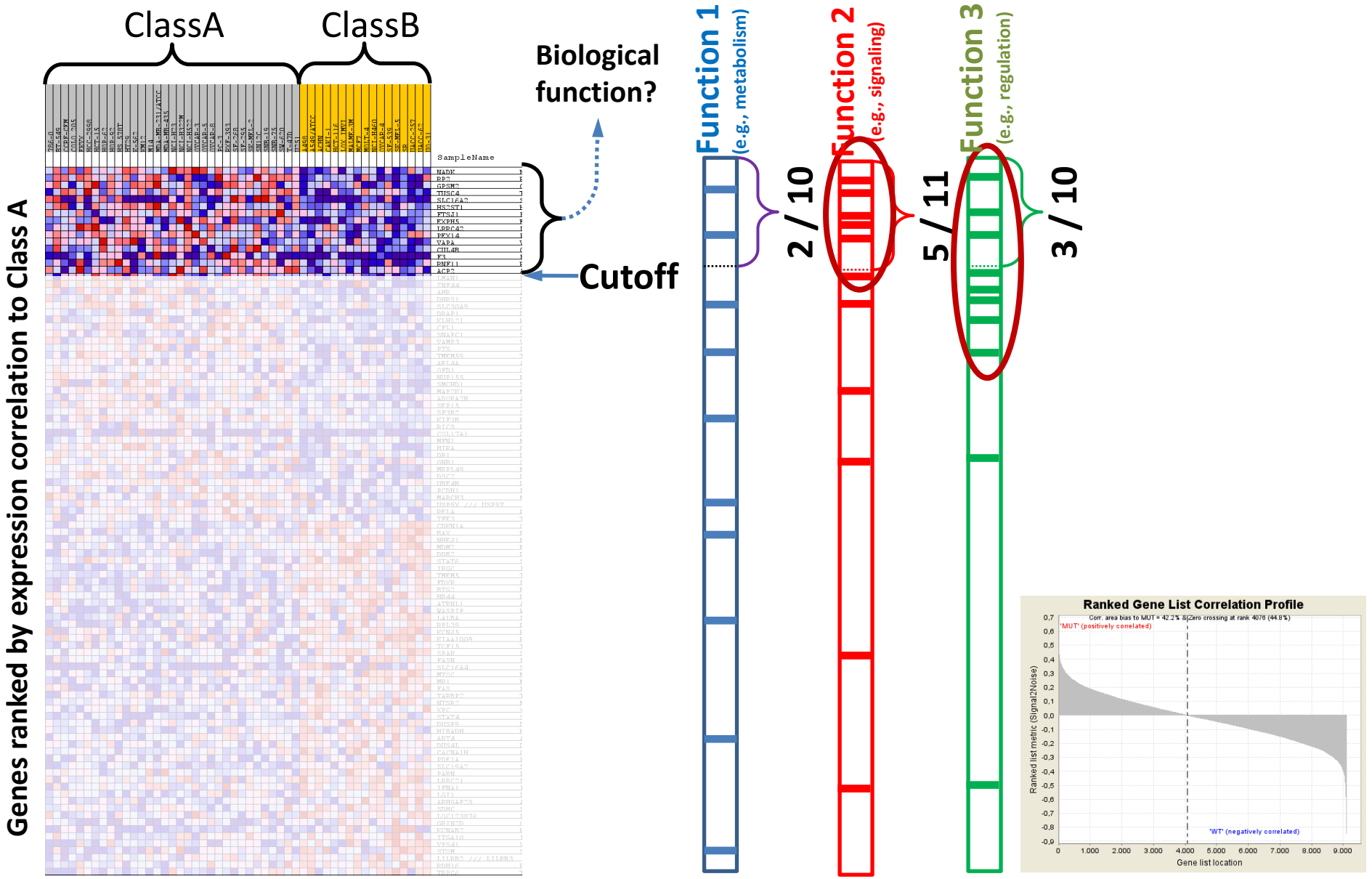
[‡]To whom correspondence may be addressed. E-mail: lander@broad.mit.edu or mesirov@broad.mit.edu.

(Subramanian et al. PNAS. 2005.)

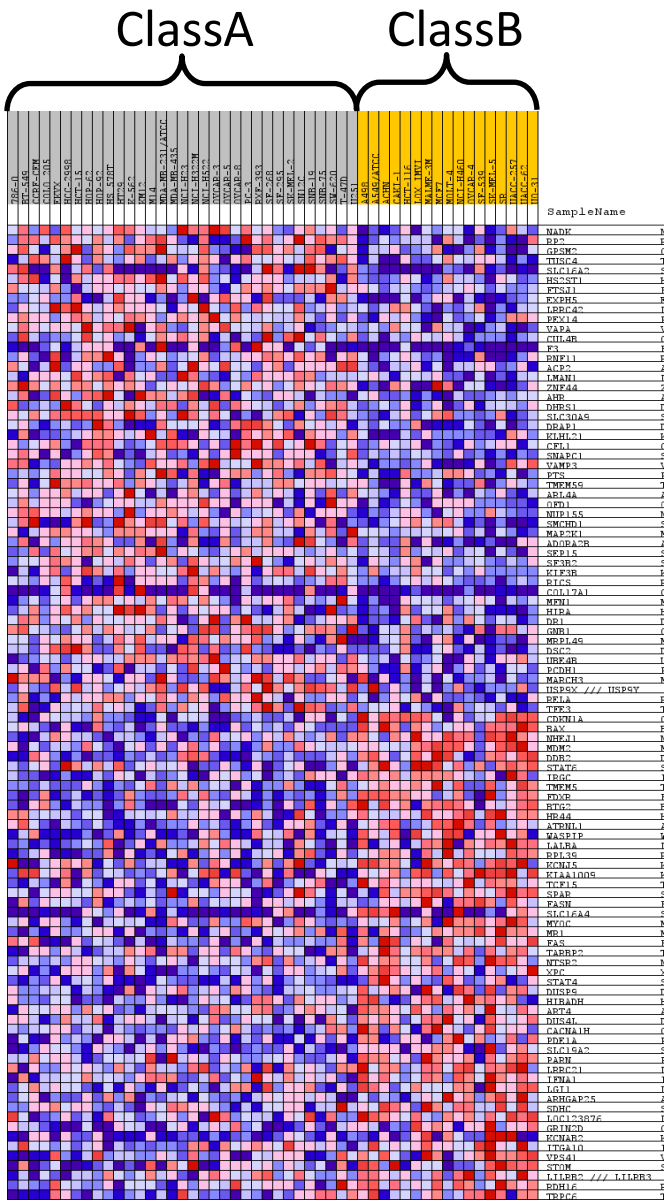
GSEA key features

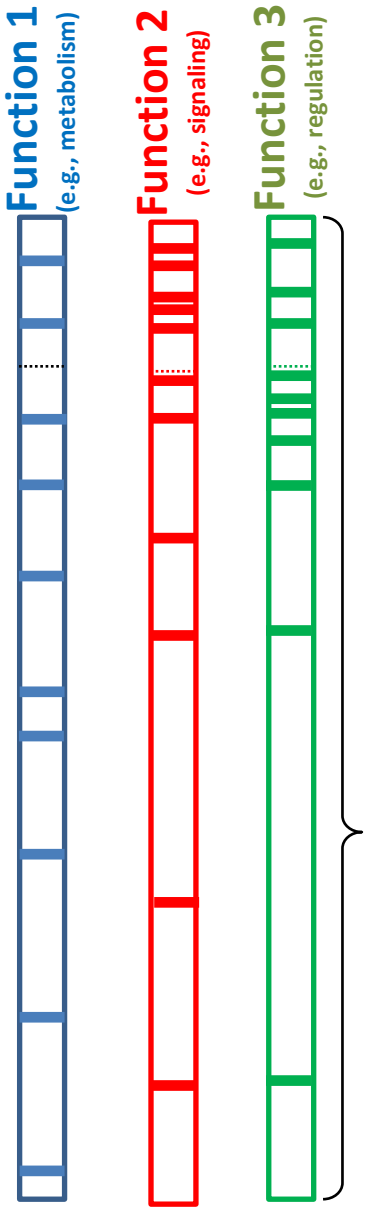
- Does not require setting a cutoff!
- Identifies the set of relevant genes as part of the analysis!
- Calculates a score for the enrichment of a **entire set of genes** rather than single genes!
- Provides a more robust statistical framework!

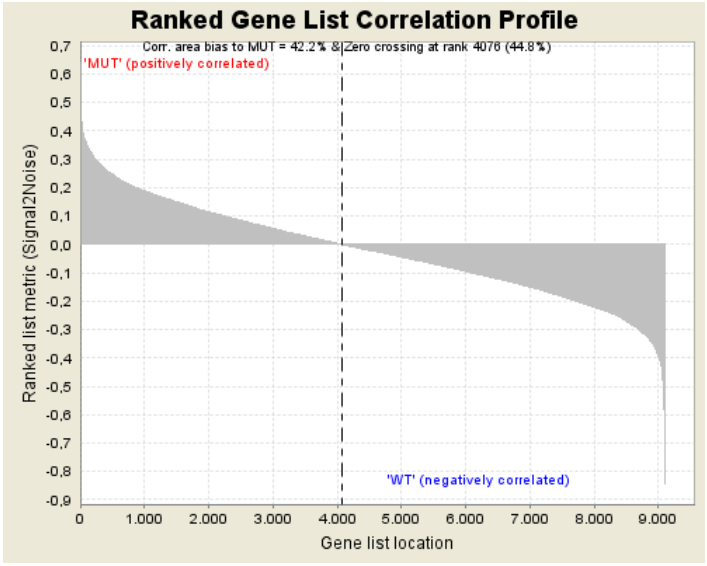
Gene Set Enrichment Analysis



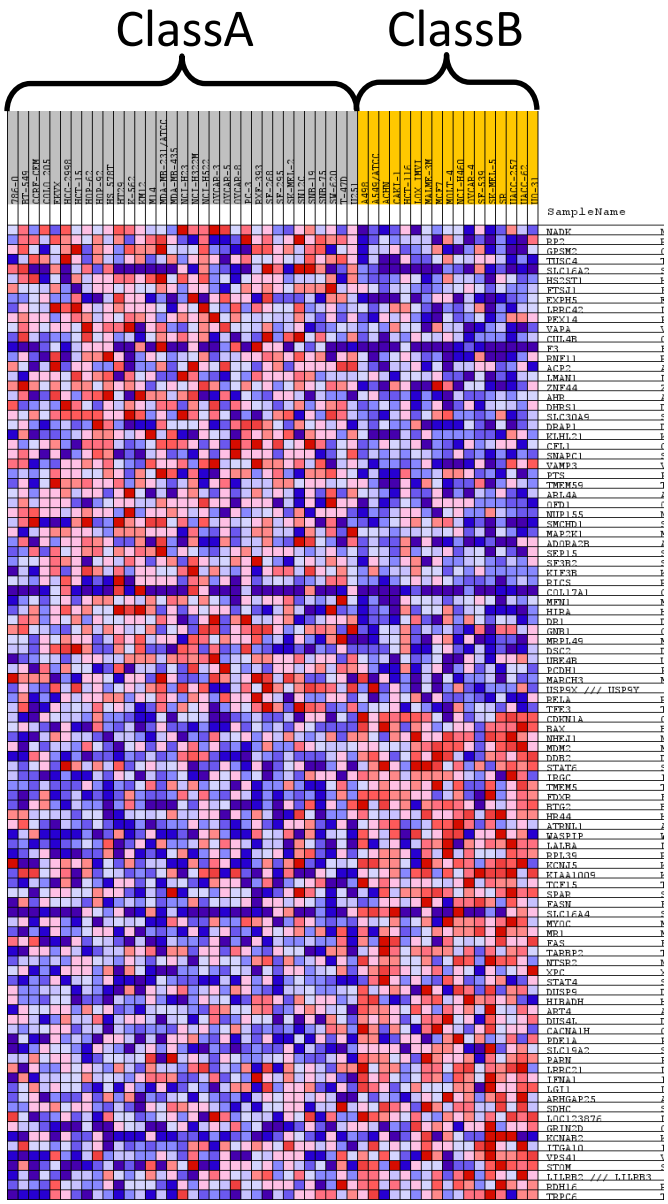
Gene Set Enrichment Analysis

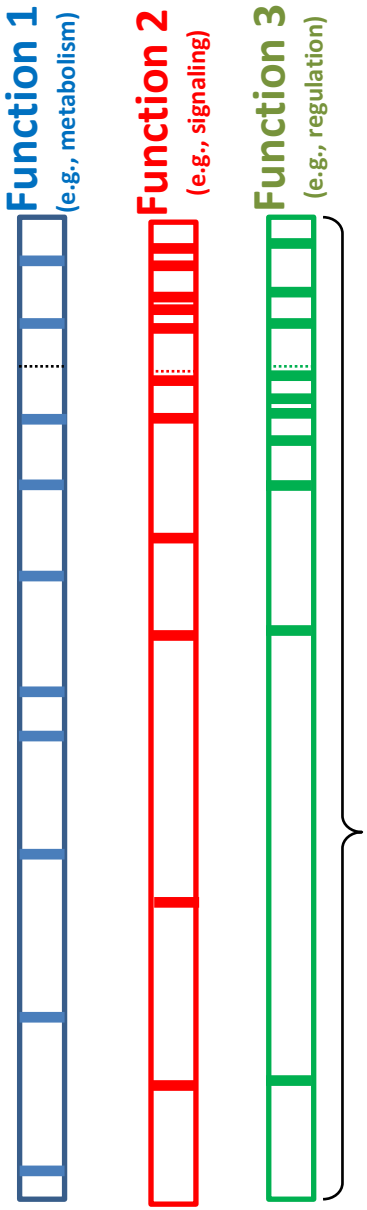


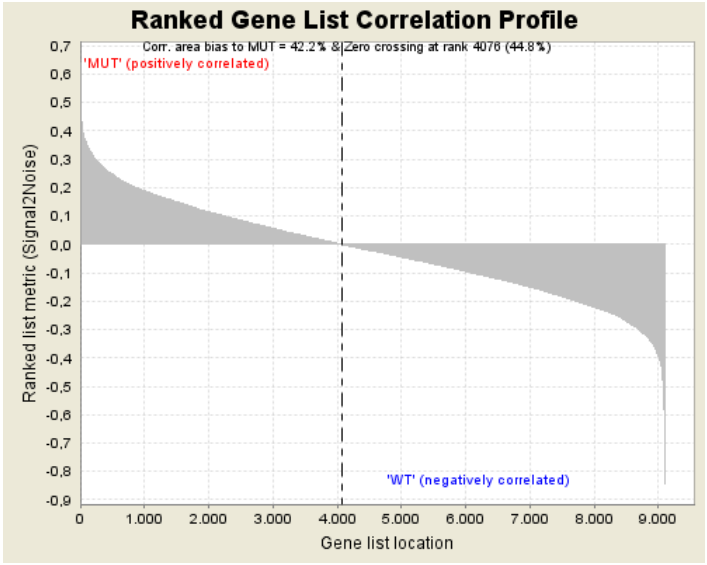


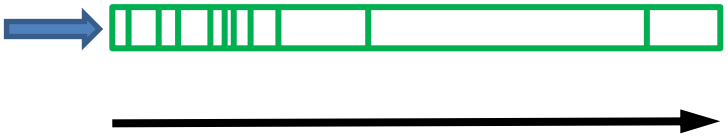


Gene Set Enrichment Analysis









Running sum:

Increase when gene annotated with the function under study

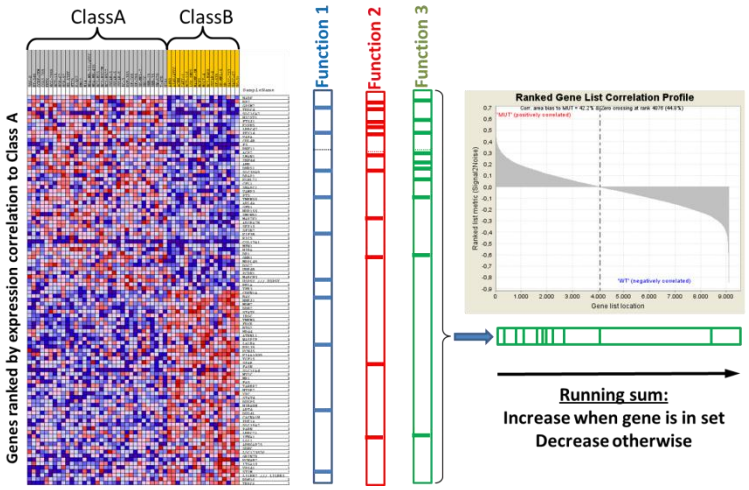
Decrease otherwise

Gene Set Enrichment Analysis

What would you expect if ALL genes annotated with this function cluster at the top of the list?

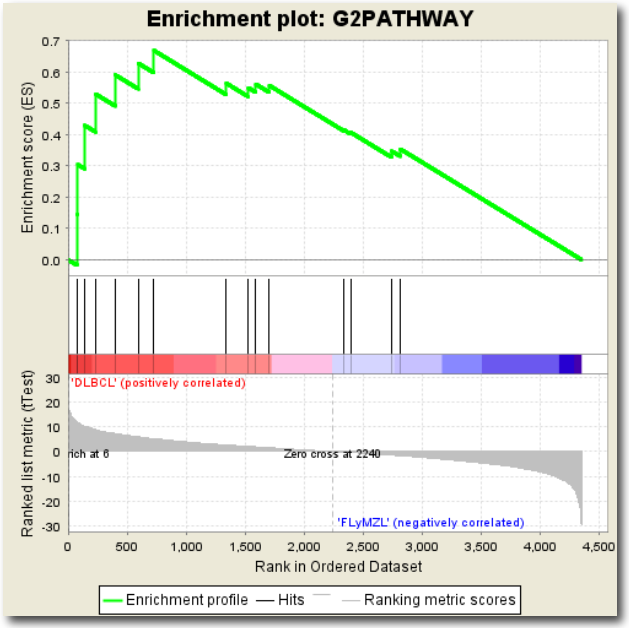
What would you expect if genes annotated with this function are randomly distributed?

What would you expect if most of the genes annotated with this function cluster at the top of the list?

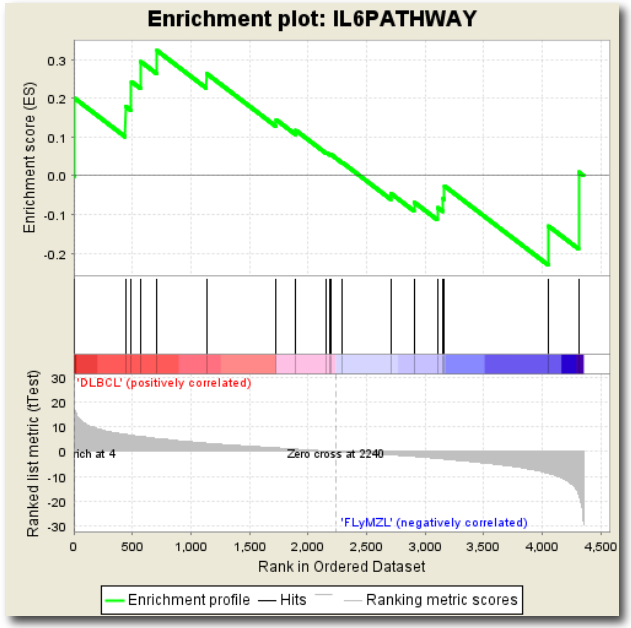


Gene Set Enrichment Analysis

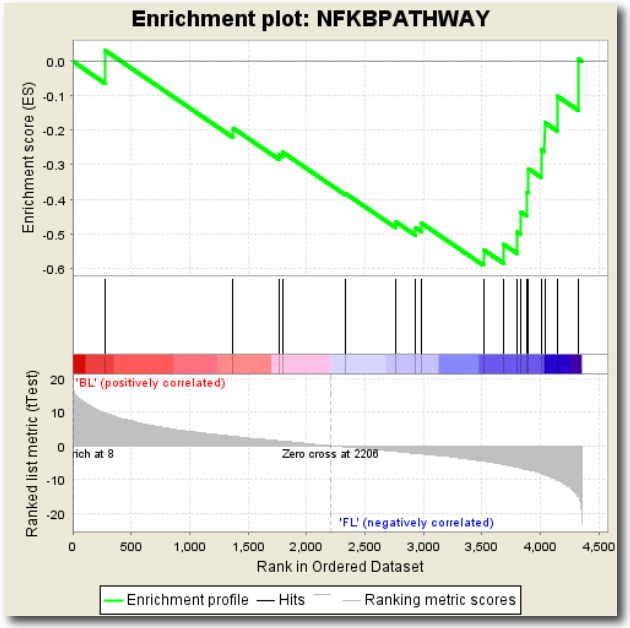
ES = 0.69

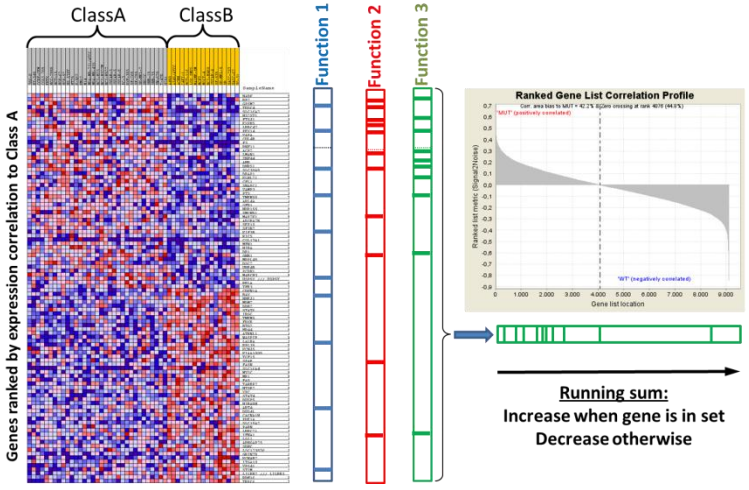


Low ES (evenly distributed)



ES = -0.59



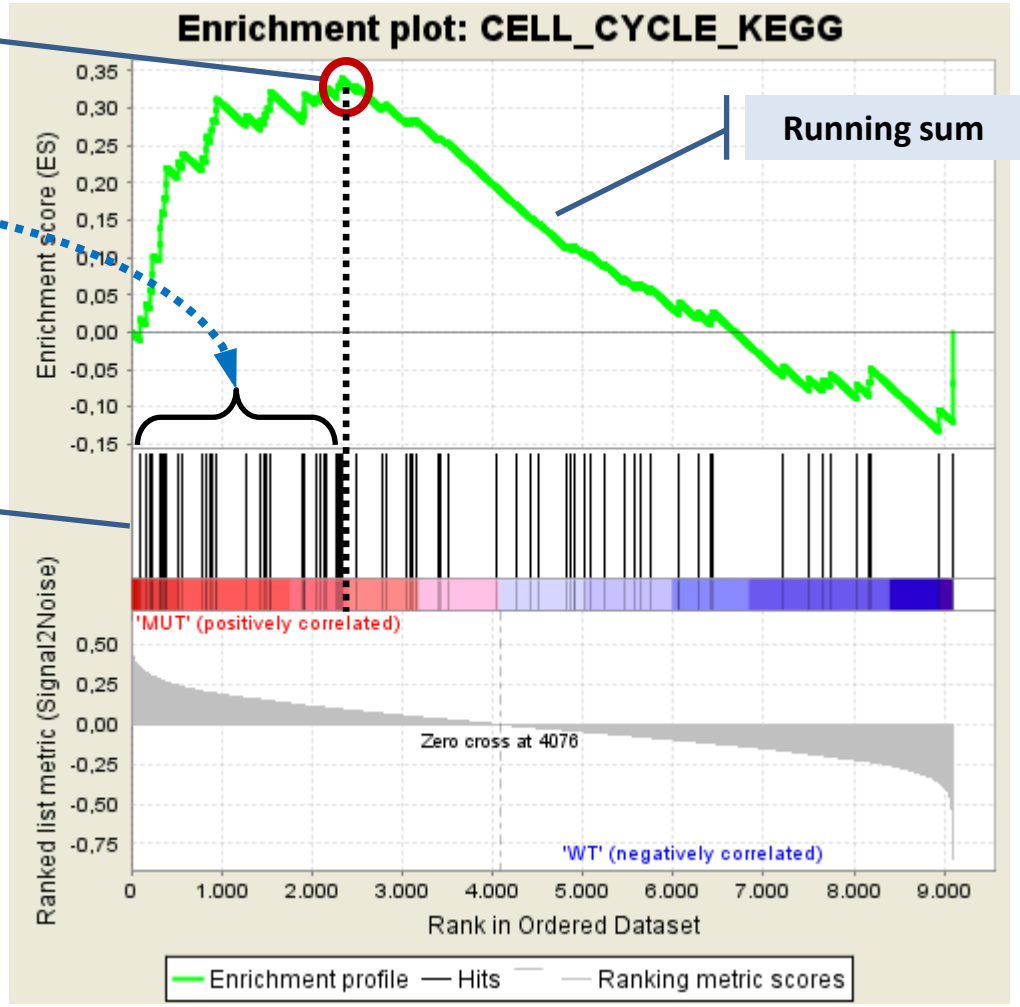


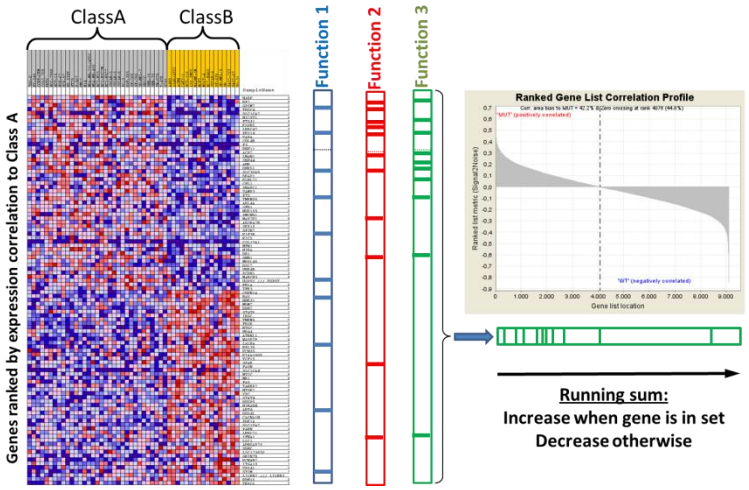
Gene Set Enrichment Analysis

Enrichment score (ES) = max deviation from 0

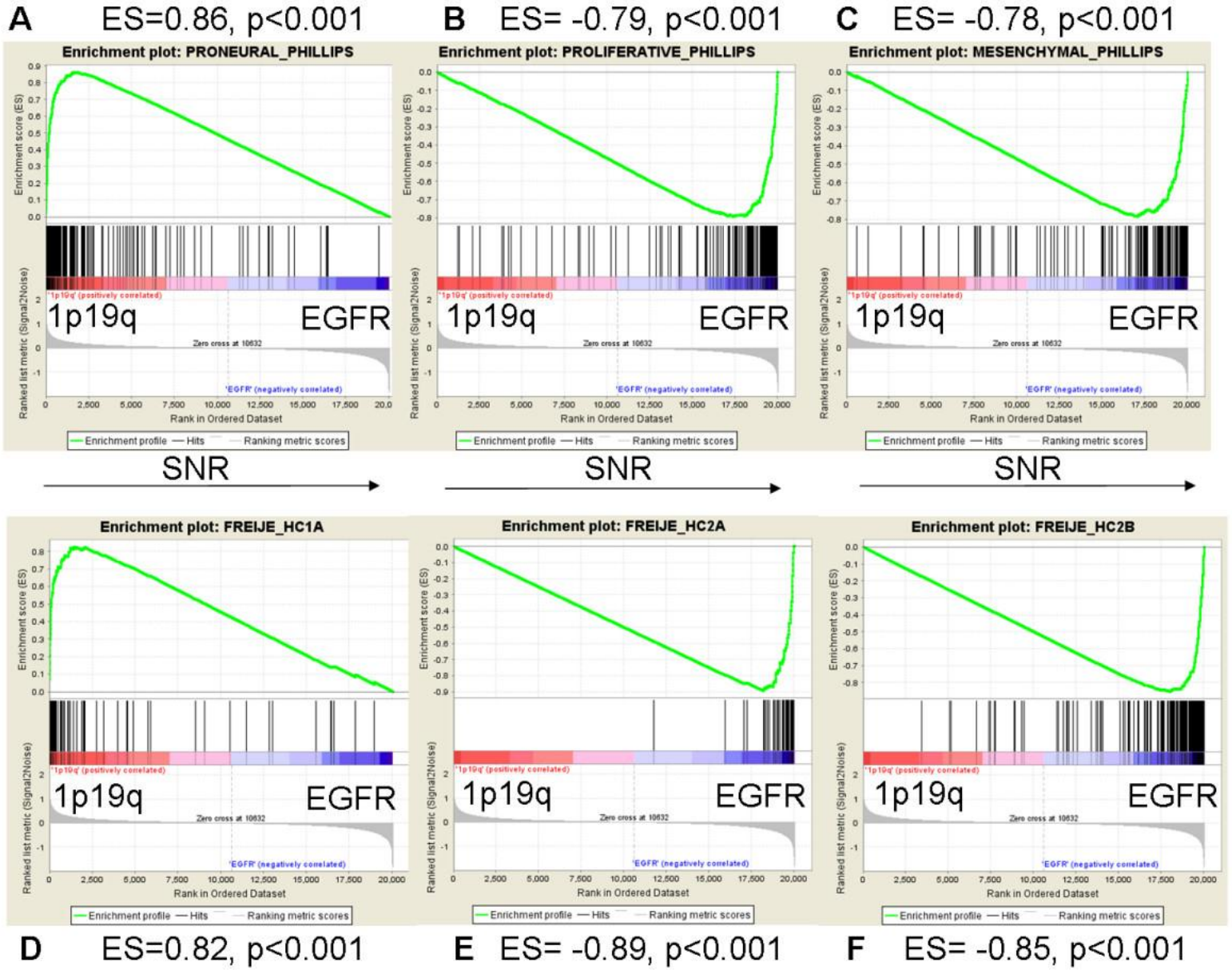
Leading Edge genes

Genes within functional set (hits)





Gene Set Enrichment Analysis

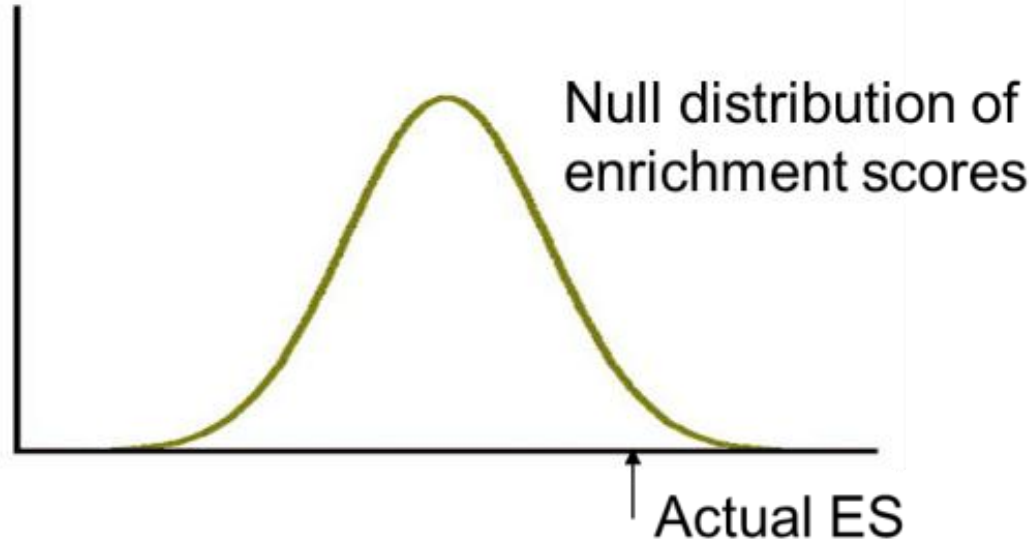


Ducray *et al. Molecular Cancer* 2008 **7**:41

Estimating Significance of ES

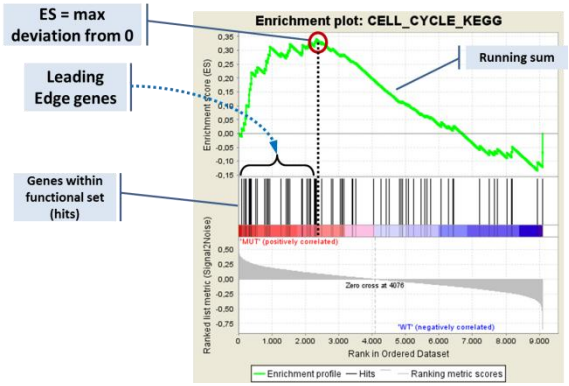
Estimating Significance of ES

- An empirical permutation test
- Phenotype labels are shuffled and the ES for this functional set is recomputed. Repeat 1000 times.
- Generating a null distribution

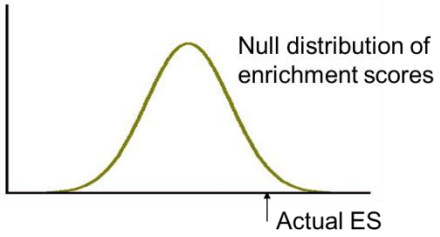


GSEA Steps

1. Calculation of an enrichment score (ES) for each functional category



2. Estimation of significance level of the ES
   - Shuffling-based null distribution



3. Adjustment for multiple hypotheses testing
   - Necessary if comparing multiple gene sets (i.e.,functions)
   - Computes FDR (false discovery rate)