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

Genes ranked by expression correlation to Class A

ClassA

ClassB

Biological function?

Cutoff

Sample vs. Class A

Ranked Gene List Correlation Profile

- MUT (positively correlated)
- WT (negatively correlated)
Enrichment Analysis

Genes ranked by expression correlation to Class A

Class A

Class B

Biological function?

Cutoff

Function 1 (e.g., metabolism)

Function 2 (e.g., signaling)

Function 3 (e.g., regulation)

2 / 10

5 / 11

3 / 10

Ranked Gene List Correlation Profile

WT (positively correlated)

WT (negatively correlated)
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

(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

Class A

Class B

Genes ranked by expression correlation to Class A

Biological function?

Cutoff

Function 1 (e.g., metabolism)

2 / 10

Function 2 (e.g., signaling)

5 / 11

Function 3 (e.g., regulation)

3 / 10

Ranked Gene List Correlation Profile

WT (positively correlated)

WT (negatively correlated)
Gene Set Enrichment Analysis

Genes ranked by expression correlation to Class A

Class A

Class B

Function 1
(e.g., metabolism)

Function 2
(e.g., signaling)

Function 3
(e.g., regulation)

Ranked Gene List Correlation Profile

Corr. gene list to MUT + 42.2% S/genes differing at least 4.578 (44.3%)
MUT (positively correlated)

WT (negatively correlated)
Gene Set Enrichment Analysis

Class A
Class B

Genes ranked by expression correlation to Class A

Running sum:
Increase when gene annotated with the function under study
Decrease otherwise

Function 1 (e.g., metabolism)
Function 2 (e.g., signaling)
Function 3 (e.g., regulation)

Ranked Gene List Correlation Profile
Can gene lists to MUT + 42.2% Sperm count at rate 47.9 (44.1%) MUT (positively correlated)
WT (negatively correlated)

Gene list location
Gene Set Enrichment Analysis

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?

What would you expect if most 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)**

**Enrichment plot: CELL_CYCLE_KEGG**

- Running sum
- Enrichment score (ES)
- Zero cross at 4076
- ‘MUT’ (positively correlated)
- ‘WT’ (negatively correlated)

**Genes ranked by expression correlation to Class A**

- Increase when gene is in set
- Decrease otherwise

**Legend**

- Enrichment profile
- Hits
- Ranking metric scores
Gene Set Enrichment Analysis

A  ES=0.86, p<0.001

B  ES=-0.79, p<0.001

C  ES=-0.78, p<0.001

D  ES=0.82, p<0.001

E  ES=-0.89, p<0.001

F  ES=-0.85, p<0.001

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

Null distribution of enrichment scores
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)