Gene Set Enrichment Analysis

Genome 559: Introduction to Statistical and Computational Genomics
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
  - GO terms are linked in a complex structure

- Enrichment analysis:
  - Find the “most” differentially expressed genes
  - Identify functional annotations that are **over-represented**
  - Modified Fisher's exact test
A quick review: Modified Fisher's exact test

Do I have a surprisingly high number of blue genes?

Null model: the 8 genes/balls are selected randomly

So, if you have 50 balls, 10 of them are blue, and you pick 8 balls randomly, what is the probability that \( k \) of them are blue?
A quick review: Modified Fisher's exact test

Hypergeometric distribution

\[ P(\sigma_t = k) = \frac{\binom{m_t}{k} \binom{m-m_t}{n-k}}{\binom{m}{n}} \]

\( m=50, \; m_t=10, \; n=8 \)

So ... do I have a surprisingly high number of blue genes?

Can such high numbers (4 or above) occur by change?

What is the probability of getting at least 4 blue genes in the null model?

\( P(\sigma_t >=4) \)
Enrichment Analysis

Genes ranked by expression correlation to Class A

ClassA

ClassB

Biological function?

Cutoff
Enrichment Analysis

Class A

Class B

Genes ranked by expression correlation to Class A

Cutoff

Biological function?

Function 1
(e.g., metabolism)

Function 2
(e.g., signaling)

Function 3
(e.g., regulation)

2 / 10

5 / 11

3 / 10

[Graph showing ranked gene list correlation profile]
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

(Subramanian et al. PNAS. 2005.)
GSEA key features

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

Class A

Class B

Biological function?

Cutoff

Genes ranked by expression correlation to Class A

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
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)

Running sum:
Increase when gene is in set
Decrease otherwise
Gene Set Enrichment Analysis

What would you expect if the hits were randomly distributed?

What would you expect if most of the hits cluster at the top of the list?
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

**Running sum:**
Increase when gene is in set
Decrease otherwise
**Gene Set Enrichment Analysis**

**ES = 0.43**
- Enrichment plot: G2PATHWAY
- Rank in Ordered Dataset
  - DLSCL (positively correlated)
  - IL6 (negatively correlated)

**ES = -0.45**
- Enrichment plot: NFkBPATHWAY
- Rank in Ordered Dataset
  - IL6 (positively correlated)

**Low ES (evenly distributed)**
- Enrichment plot: IL6PATHWAY
- Rank in Ordered Dataset
  - DLSCL (positively correlated)
  - IL6 (negatively correlated)

**Classes**
- Class A
- Class B

**Ranked Gene List Correlation Profile**
- Increase when gene is in set
- Decrease otherwise
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
1. Calculation of an enrichment score (ES) for each functional category

2. Estimation of significance level of the 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

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