Gene Set Enrichment Analysis

Isaac Blackhurst

References

- Mootha et al. Nature Genetics 34(3):267-273 (2003)
- Subramanian et al. PNAS 102(43):15545-15550 (2005)

Purpose

Determine if a set of genes is differentially expressed between two conditions

Like the Global Test, the set tested in GSEA should be biologically determined

Individual gene t-tests

A few problems we run into when we test individual genes

- No genes are significant after correcting for multiple hypotheses
- Lots are significant, but no biological story as to why
- Lots of genes with a similar purpose changing a little might be more important that one gene changing a lot
- There is too little overlap of significant genes in different studies of the same system

Overview of GSEA

- Rank the genes according to some measure of correlation with the condition
- Calculate and enrichment score for a gene set of interest
- Take the maximal enrichment score for that set
- Permute the condition labels to get a null distribution of enrichment scores
- Get the p-value by looking at the percent of enrichment scores that are more extreme than the one under the true condition labels
- If this is done for many sets, make an adjustment for multiple hypothesis

Notation

A and B are our clinical outcomes
N-sub S is the number of genes a gene set
N is the total number of genes
r-sub i is the within gene correlation with the clinical outcome
p is a weight parameter

Rank the genes in a list L

Rank genes according to some metric of correlation
The authors use signal to noise ratio, which is

$$\frac{\mu_A - \mu_B}{s_A + s_B}$$

where A and B are the two conditions
Also can use a t-test or the correlation coefficient

Calculate the Enrichment Score (ES)

Method I (Original method)

Create a running sum statistic based on the following
If gene p is not in set S, then add

$$X_i = -\sqrt{\frac{N_S}{N - N_S}}$$

If gene p is in set S, then add

$$X_i = \sqrt{\frac{N - N_s}{N_S}}$$

This creates a running sum
The maximum sum over the whole list L is the Enrichment Score MES

Image of process

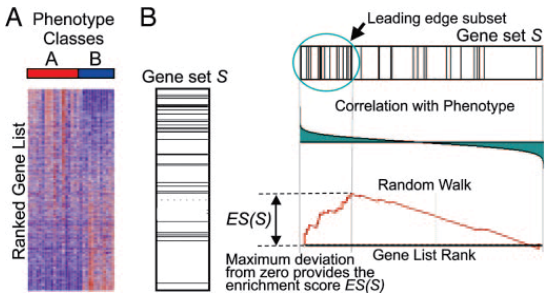

Image from Subramanian et al.

Calculate the Enrichment Score (ES)

Method II

Create a running sum statistic based on the following
If gene p is not in set S, then subtract

$$X_i = \frac{1}{N - N_s}$$

If gene p is in set S, then add

$$X_i = \frac{|r_i|^p}{N_R} \quad \text{where} \quad N_R = \sum_{i \in S} |r_i|^p$$

This creates a running sum
The maximum sum over the whole list L is the Enrichment Score
MES

Enrichment Score

- If the genes in set S are randomly distributed throughout the list L, then the score should never be very high
- If they are concentrated at the top or bottom of the list, it will be high
- p is a tuning parameter
- when p=0, MES is a standard Kolmogorov-Smirnov statistic
- The authors use p=1 in their paper

Examples of the process

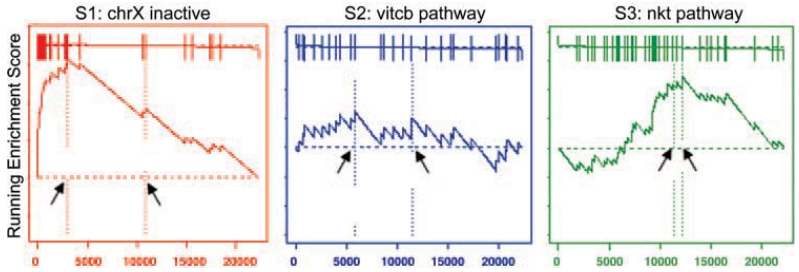

Image from Subramanian et al.

Estimation of Significance

- Under the null hypothesis, the genes are randomly distributed throughout the list L
- To get a null distribution, we permute the labels A and B, our clinical outcomes, and recalculate the MES for a large number of permutations, say 1000
- This creates a histogram of MES
- The p-value is created by looking at the proportion of MES that are more extreme than the MES using the true labels

Multiple Hypothesis Correction

- When many different gene sets are tested, we need to correct for multiple hypothesis testing
- Normalize the MES to account for the size of the gene set to create a normalized statistic NES
- Control the proportion of false positives by controlling the false discovery rate
- Authors used Benjamini-Yekutieli

Advantage over other tests of set

Leading edge subset that consists of those groups of genes that are responsible for the enrichment score

R packages

- This can be done in R using a number of packages
- seqGSEA
- GSEAlm
- PGSEA

Summary

Order the genes by correlation with clinical outcome
Compute the enrichment score by taking the maximum running sum total
Permute the clinical outcomes to generate a distribution under the null that the genes in the set are randomly distributed
Correct for multiple hypothesis testing

Questions?