

Gene Set Enrichment Analysis

Subramanian et. al. 2005

Motivation

Goal: Determine which genes have significant expression change under a condition

Typical Analysis: Choose a threshold of expression difference

Motivation: Problems

No genes may be significantly altered

Lots of noise

-or-

Many significantly altered genes

Hard to interpret, probably noise

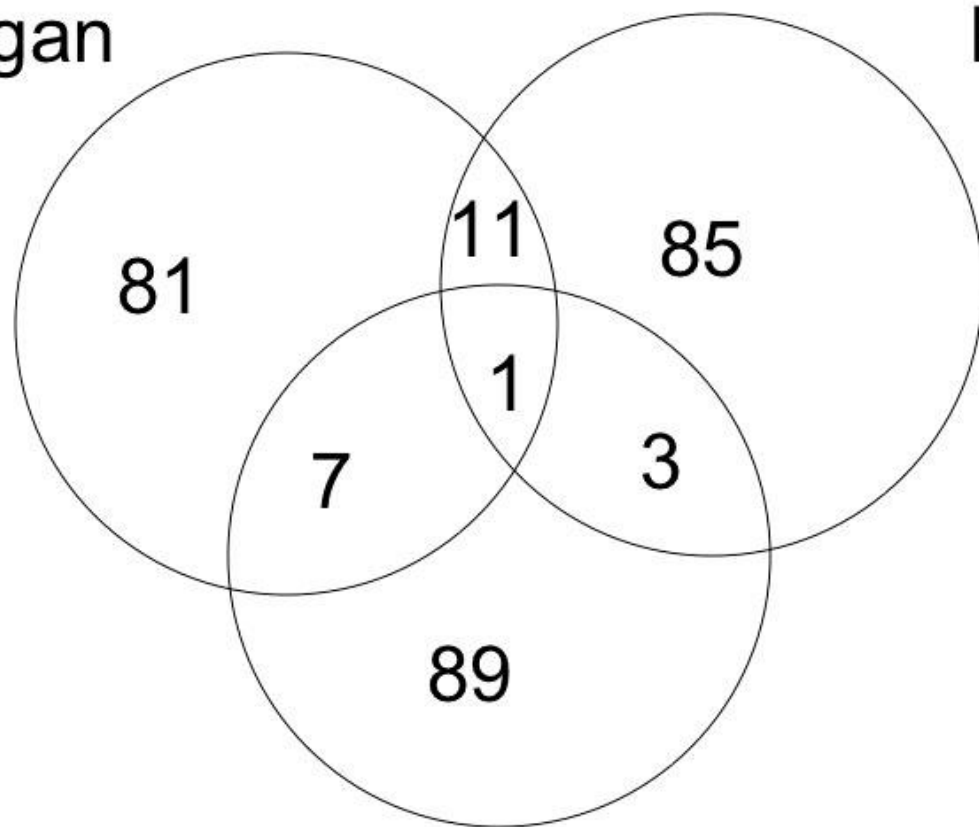
Motivation: More Problems

Misses cumulative effects from many slightly altered genes

“An increase of 20% across all genes encoding members of a metabolic pathway...may be more important than a 20-fold increase in a single gene”

Michigan

Boston



Stanford

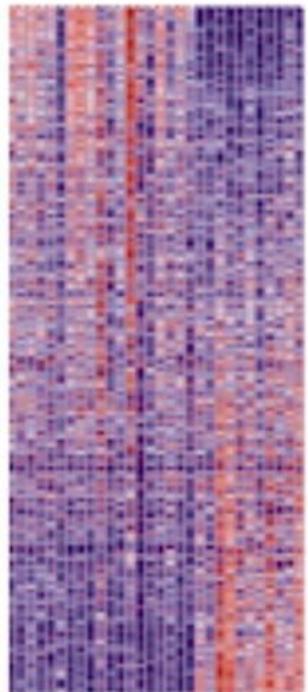
GSEA: The basics

Gene Set Enrichment Analysis

Solves problems by using sets of genes

Sets come from prior biological knowledge

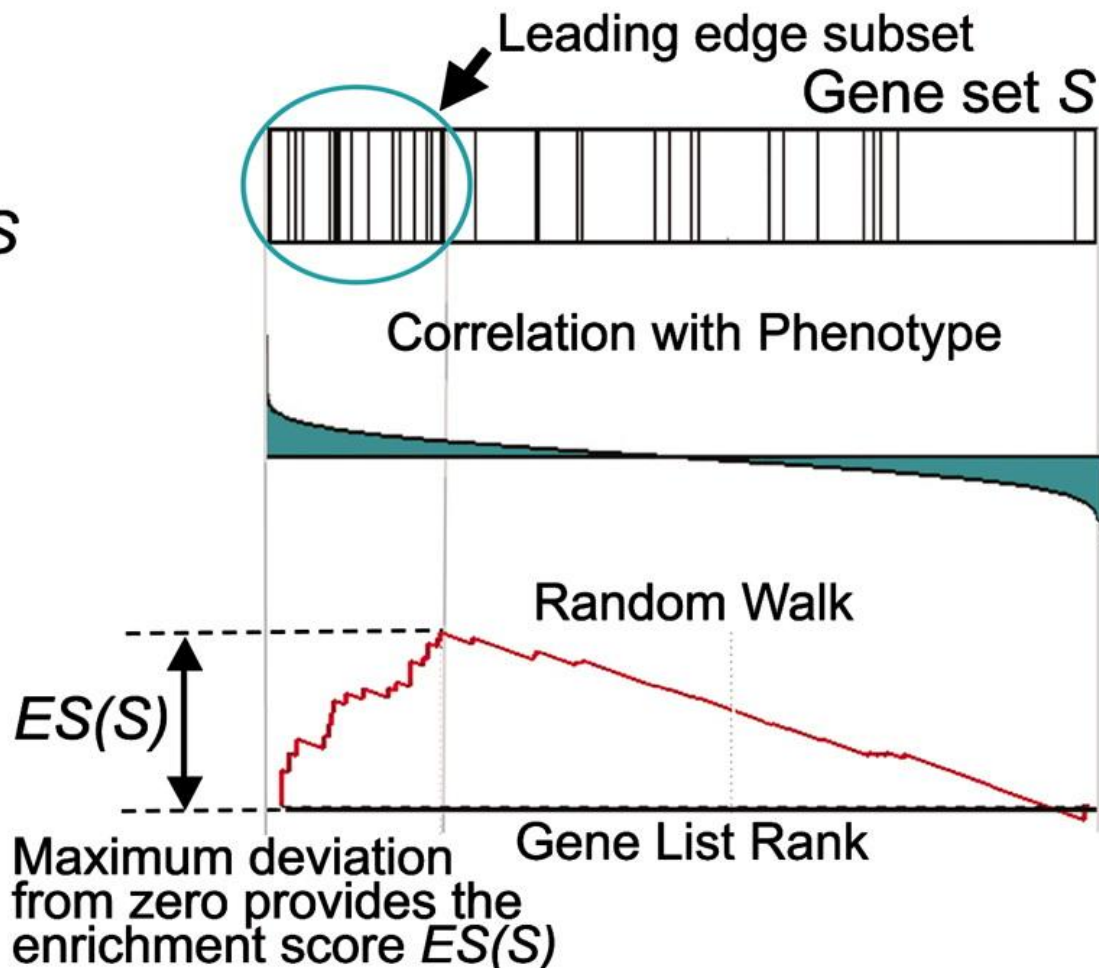
A Phenotype
Classes
A B



Ranked Gene List

B

Gene set S



GSEA: Basics

Given: a set **S** of genes and a list **L** of genes ranked by correlation (or other metric) between two conditions/classes/phenotypes

Question: is **S** randomly distributed in **L** or is **S** focused at one of the ends?

GSEA: Details

Calculating Enrichment Score (ES):

For all positions i in \mathbf{L} (p is a parameter)

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

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

Find the largest (inc. negative) value for $P_{\text{hit}} - P_{\text{miss}}$

GSEA: Details

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

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

When p is 0, this is the fraction of genes in S versus not in S up until point i

(This case happens to correspond to the Kolmogorov-Smirnov statistic)

(if you don't know what that is don't worry about it)

GSEA: Getting the Significance

Randomly reassign class labels and recompute the ES 1000 times

Compute P-value of the observed ES by comparing it to the distribution of ES scores

If performing with multiple candidate sets correct with FDR

Analyzing GSEA

Leading Edge Subset - the subset of genes in the set S which appear before the max ES value

GSEA can also be used for multiple sets and alternate rankings

MSig DB

The unintentional star of the paper:

The hand curated database of gene sets from which **S** is chosen

Contains 1,325 gene sets in 4 collections in V1.
0

MutSig DB

Still Updated Today: [Link](#)

Now contains 10348 sets in 8 collections for V5.0

Used in a large variety of studies

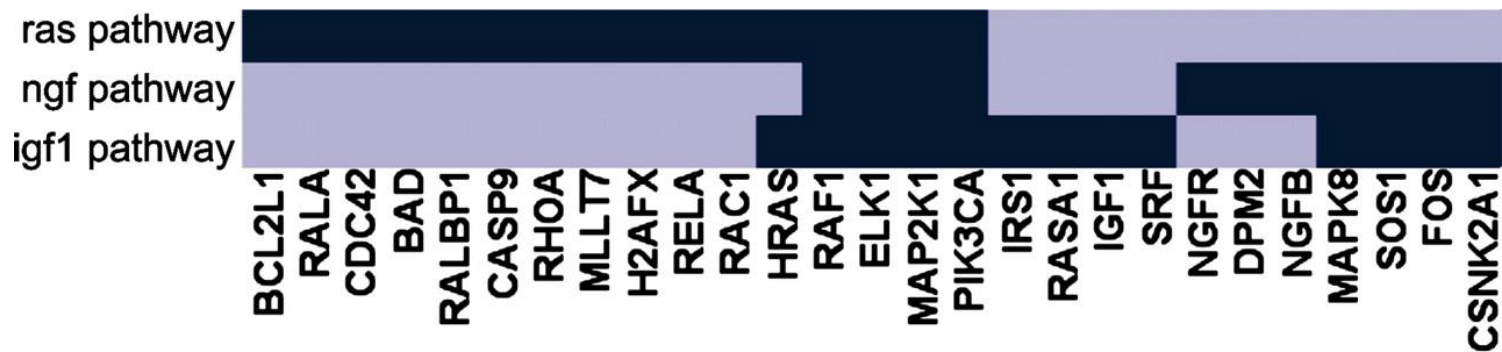
Results: Proof of Concept

Dataset of 15 male and 17 female lymphoblastoid cell lines

Looked at phenotypes “male>female” and “female>male”

Found mostly Y chromosome sets for male > female, and reproductive tissue gene sets

Results: p53 In Cell Lines

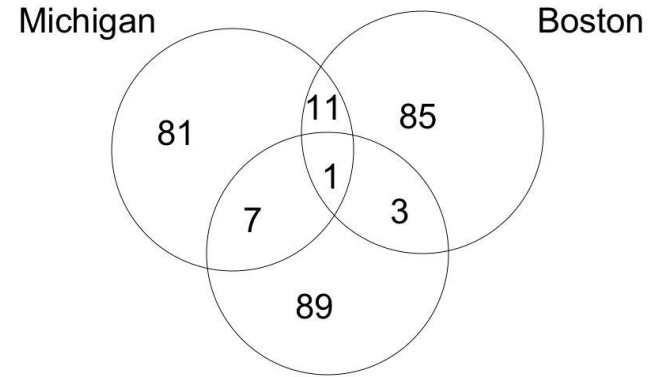


Results: Lung Cancer

Michigan and Boston Studies

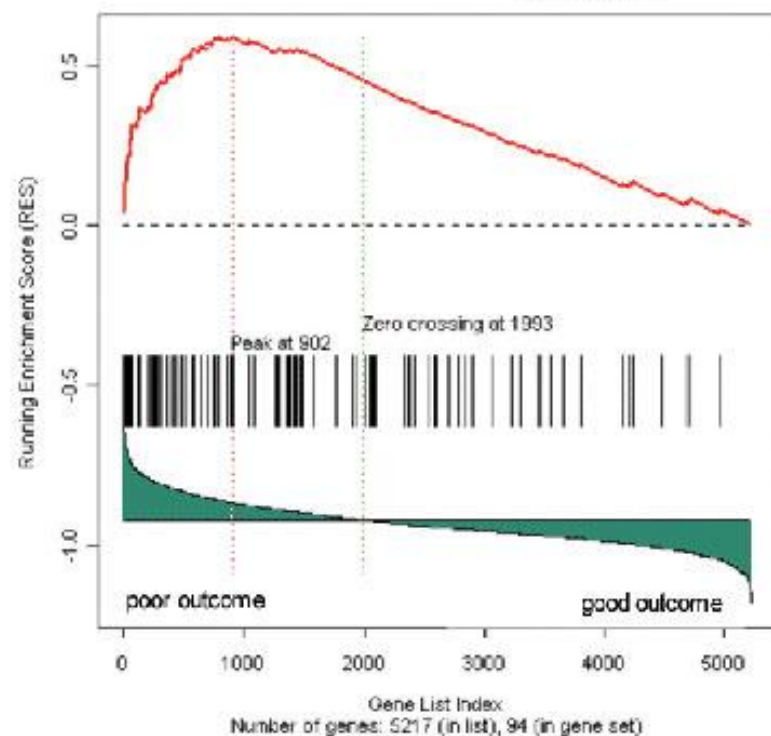
No genes were significantly associated with cancer outcome

However, GSEA found approx. half overlapping gene sets (5 of 8 to 6 of 11)



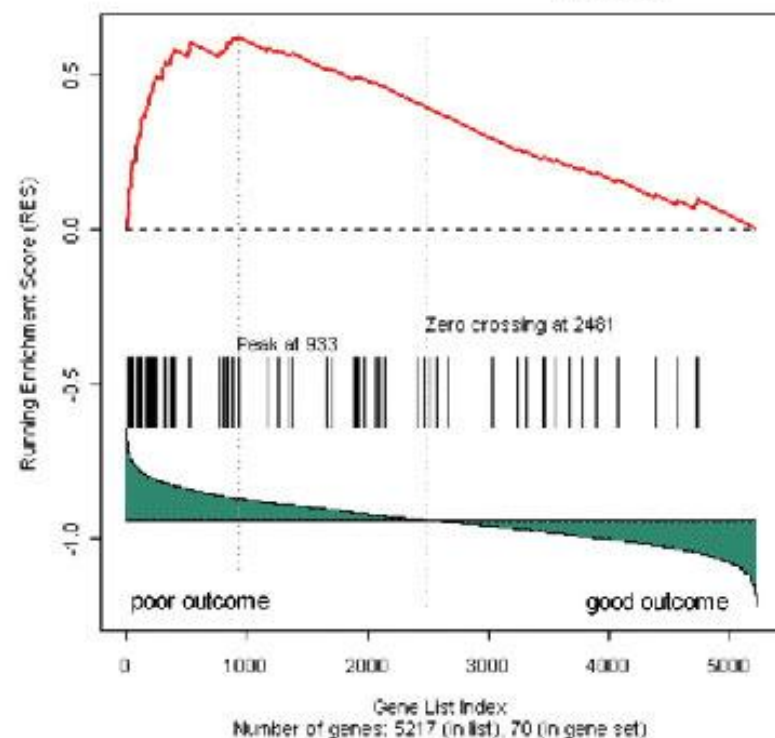
Boston Dataset

Gene Set: S_{Michigan}



Michigan Dataset

Gene Set: S_{Boston}

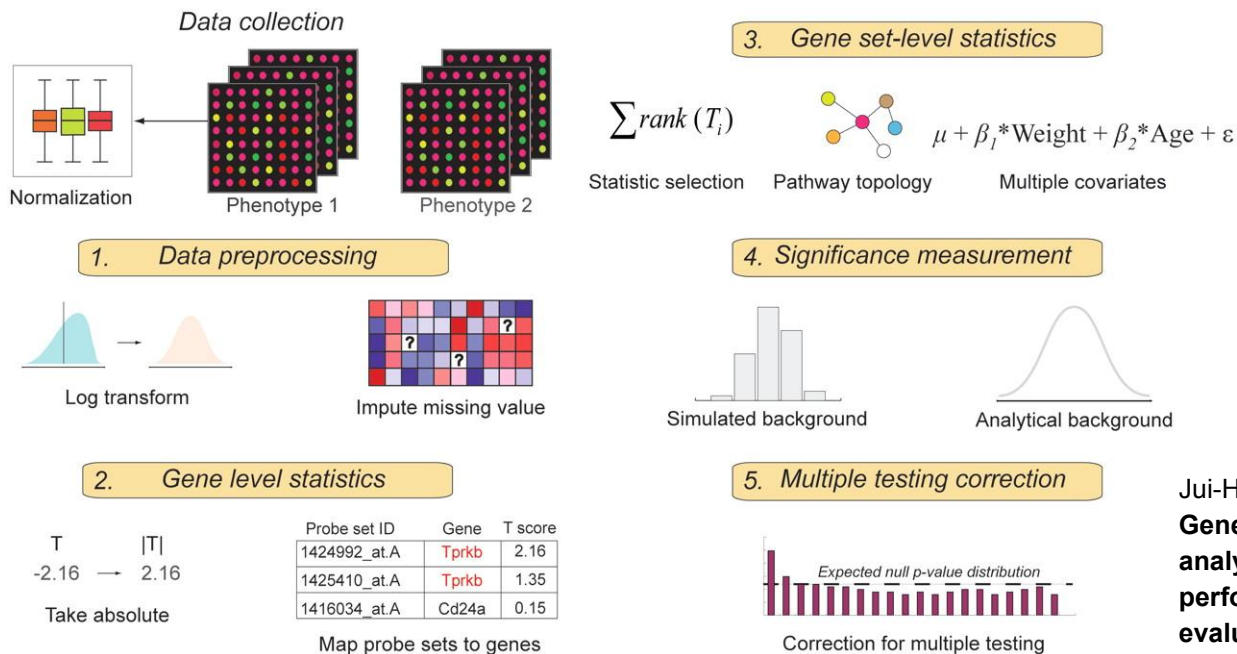


Gene set	FDR	Gene set	FDR
Data set: Lymphoblast cell lines		Data set: Lung cancer outcome, Boston study	
Enriched in males		Enriched in poor outcome	
chrY	<0.001	Hypoxia and p53 in the cardiovascular system	0.050
chrYp11	<0.001	Aminoacyl tRNA biosynthesis	0.144
chrYq11	<0.001	Insulin upregulated genes	0.118
Testis expressed genes	0.012	tRNA synthetases	0.157
Enriched in females		Leucine deprivation down-regulated genes	0.144
X inactivation genes	<0.001	Telomerase up-regulated genes	0.128
Female reproductive tissue expressed genes	0.045	Glutamine deprivation down-regulated genes	0.146
Data set: p53 status in NCI-60 cell lines		Cell cycle checkpoint	0.216
Enriched in p53 mutant		Data set: Lung cancer outcome, Michigan study	
Ras signaling pathway	0.171	Enriched in poor outcome	
Enriched in p53 wild type		Glycolysis gluconeogenesis	0.006
Hypoxia and p53 in the cardiovascular system	<0.001	vegf pathway	0.028
Stress induction of HSP regulation	<0.001	Insulin up-regulated genes	0.147
p53 signaling pathway	<0.001	Insulin signalling	0.170
p53 up-regulated genes	0.013	Telomerase up-regulated genes	0.188
Radiation sensitivity genes	0.078	Glutamate metabolism	0.200
Data set: Acute leukemias		Ceramide pathway	0.204
Enriched in ALL		p53 signalling	0.179
chr6q21	0.011	tRNA synthetases	0.225
chr5q31	0.046	Breast cancer estrogen signalling	0.250
chr13q14	0.057	Aminoacyl tRNA biosynthesis	0.229
chr14q32	0.082		
chr17q23	0.071		

Critique And Other Methods

“Surprisingly, GSEA is based on the Kolmogorov–Smirnov (K–S) test, which is well known for its lack of sensitivity and limited practical use.”

– Rafael A. Irizarry et al, Gene Set Enrichment Analysis Made Simple



Jui-Hung Hung et al.
**Gene set enrichment analysis:
performance
evaluation and usage
guidelines**