Newton *et al.*'s "Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis"

Annals of Applied Statistics (2007), 1, 85-106.

Presented by Fred Boehm

Statistics 992
1 April 2013

Table of Contents

Presentation goals

1. Describe *Newton et al.*'s flexible approach to gene set enrichment based on random sets
2. Compare empirical & theoretical properties of random set methods with those of SAFE/GSEA

Analysis of gene expression microarray study

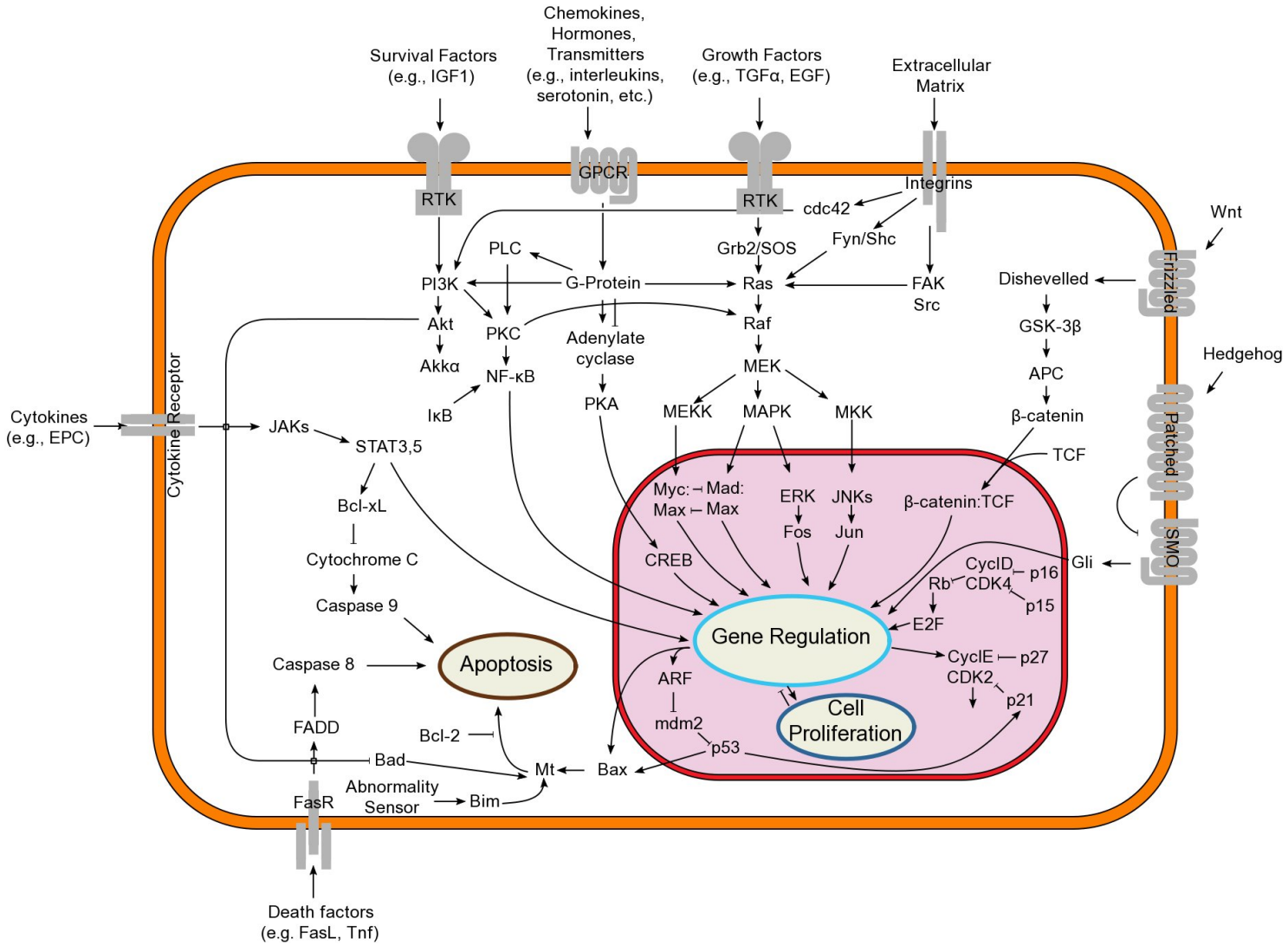
How to extract biological information from microarray results

- Identify differentially expressed genes among, for example, two classes of subjects
- Assess for related biological functions of gene products
- Gene set enrichment can be useful to identify shared biology among differentially expressed genes

Definitions

- Gene set: a collection of genes whose products are known to share biological function
  - Examples include genes whose products participate in a single known cellular signaling cascade
  - For present purposes, we'll focus on gene sets in the Gene Ontology database
- Gene set enrichment: over-representation of differential expression signal in a given gene set

Cell signaling pathways as examples of GO sets



Two existing approaches to gene set enrichment

Selection

- Choose a short list of genes with 'most altered' expression levels
- Evaluate, via Fisher's exact test (or similar test) intersection of short list and functional GO sets to get a score per GO set
- A GO set score is high if far more than expected short list genes are in the GO set

SAFE/GSEA permutation

- Retain information on all genes & permute gene labels to measure significance of set-level statistics from gene-level statistics

Limitations of above approaches

Limitations of selection approach

- Enrichment results depend on selection stringency
- Gives equal weight to genes at both 'ends' of the short list

Limitations of SAFE/GSEA permutation approach

- Computational burden, since it uses microarray data themselves, rather than results of DE analysis

Overview of *Newton et al.*'s approach

- Borrows from both SAFE/GSEA and selection approaches to combine GO set-level statistics (like SAFE/GSEA) but calibrate them like Fisher's exact test calibrates the intersection of a functional GO set and a short list of genes
- Calibration is conditional on DE analysis results since *Newton et al.* consider set-level statistics that would be achieved by random sets of genes.
- *Newton et al.* derive formulas for mean and variance of this conditional distribution of gene set scores
  - Hence, Monte Carlo methods may not be needed

Random-set enrichment scoring

- Let $g \in G$ index the genes
- Denote by $\{s_g\}$ the collection of scores for the genes
  - s_g could be indicator of being on the short list of DE genes
  - Alternatively, could be a more quantitative statistic

Consider a single category C containing m genes

- Consider unstandardized enrichment scores $\bar{X} = \frac{1}{m}\sum_{g \in C} s_g$ as random variables
  - Randomness arises from nature of assignment of genes to be in C
  - Recall that we want to compare the observed gene set scores to those we would see for hypothetical sets
- Treat C as though it were drawn uniformly at random (without replacement) from the $\binom{G}{m}$ possible sets

Extension to more general quantitative gene scores

- $\bar{X}$'s distribution becomes intractable when we consider more general quantitative scores, but we can avoid MC methods with formulas for the first two moments
- Conditional on gene-level scores,

$$\mathbb{E}\bar{X} = \frac{\sum_{g=1}^{G} s_g}{G} \tag{1}$$

$$var(\bar{X}) = \frac{1}{m}\left(\frac{G-m}{G-1}\right)\left\{\left(\frac{\%sum_{g=1}^{G} s_g^2}{G}\right) - \left(\frac{\sum_{g=1}^{G} s_g}{G}\right)^2\right\} \tag{2}$$

- Consider $Z = \frac{\bar{X}-\mathbb{E}\bar{X}}{\sigma}$
  - Z has mean zero & variance 1 under H_0: C is not enriched in DE genes

Random-set enrichment scoring

Represent

Gene	1	2	3	⋯			⋯	G	
Selected	0	1	0	⋯	1	1	⋯	1	n
In category	1	1	1	⋯	1	0	⋯	0	m
In both		1			1				x

		Selected: yes	Selected: no	
C		x	$m-x$	m
not C		$n-x$	$G-n-m+x$	$G-m$
		n	$G-n$	G

Permute

Permuted	1	0	1	⋯	0	0	⋯	0	n
In category	1	1	1	⋯	1	0	⋯	0	m
In both	1		1						X

implies $X \sim$ Hypergeometric

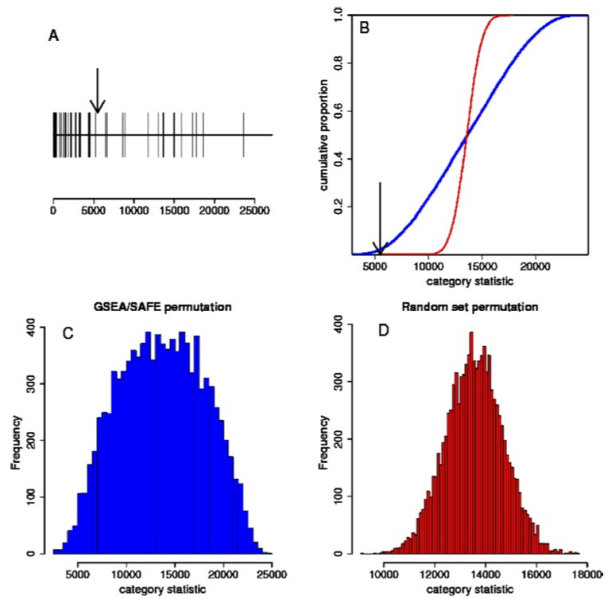
Generalize

Gene score	s_1	s_2	s_3	⋯			⋯	s_G	
In category	0	1	0	⋯	1	0	⋯	0	m
Combined		s_2			s_g				x

permuting, $X/m \sim (\mu, \sigma^2)$

- Proceed from first table to second by permuting (entries in) either of the two rows
- Then generalize to quantitative s_g, where we can still calculate mean & variance

Random sets v. SAFE/GSEA



- Panel A: rank plot of probe set correlation scores
  - $m = 48$ probe sets for a single GO category, GO:0019883
  - arrow marks the mean rank
- Random sets method shuffles the labels that are already in GO:0019883
- SAFE/GSEA shuffles labels on original chip data
- Category statistic, for this example, is rank of correlation scores, but we could use other category statistics
- SAFE p-value: 0.02; random sets p-value: $< 10^{-10}$

Two strategies with random sets

Strategy 1: Selection

- Start with a short list of extremely altered genes
- Ask if there is over-representation of in a GO category

Strategy 2: Averaging

- Averages gene-level evidence across all genes in the GO category

Comparing selection to averaging

- Each approach has a domain of superiority; neither is always preferred

Statistics

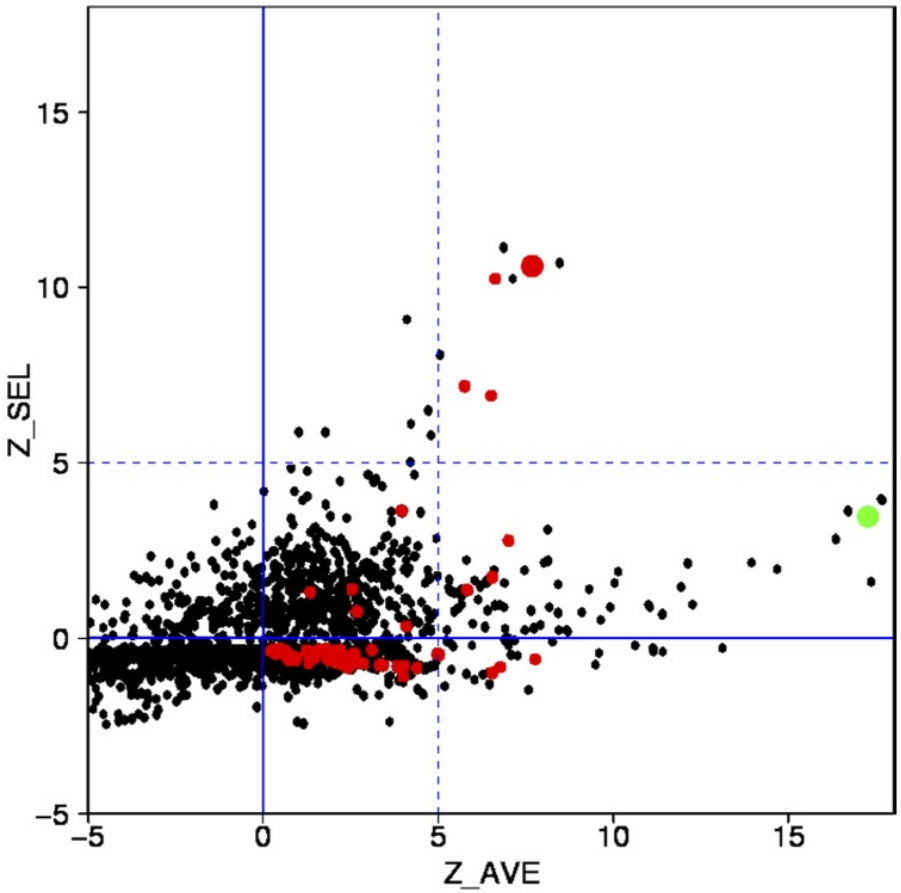
$$\bar{X}_{ave} = \frac{1}{m}\sum_{g \in C} s_g, \bar{X}_{sel} = \frac{1}{m}\sum_{g \in C} \mathbb{1}_{[s_g > k]} \tag{3}$$

Comparing GO categories

We frame the problem as a test of the null hypothesis that C is not enriched.

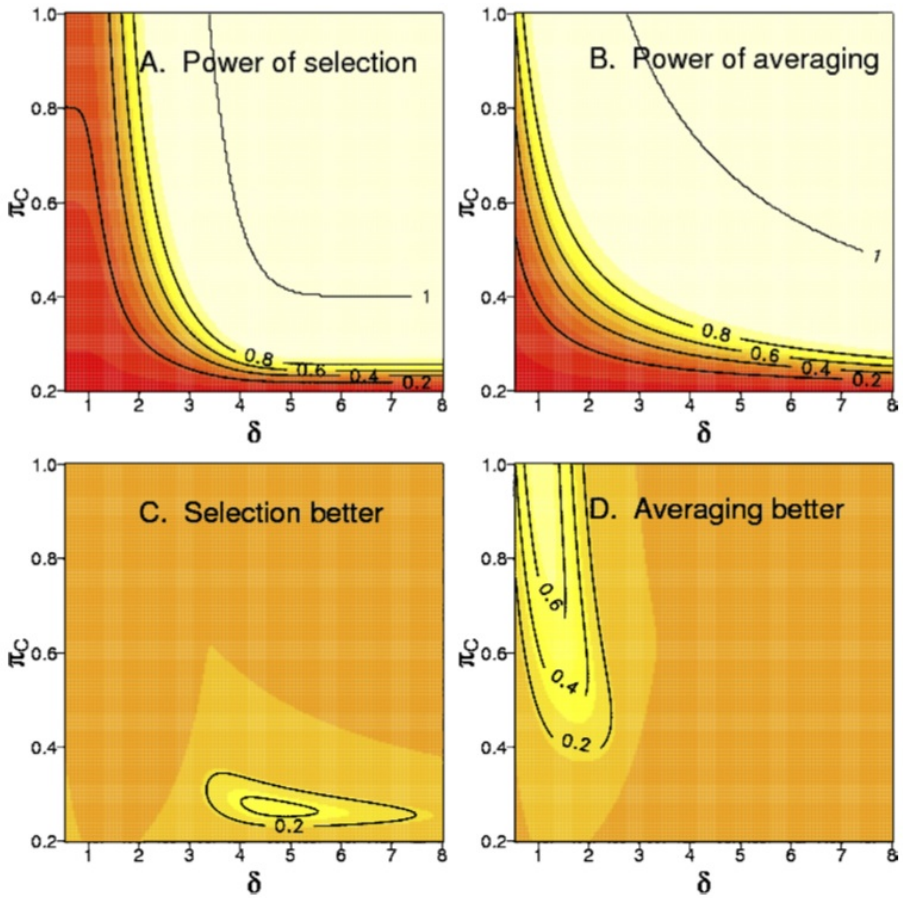
- Suppose that each gene g is either truly DE ($I_g = 1$) or not ($I_g = 0$) between two states
- Let $\pi = \frac{1}{G} \sum_{g=1}^{G} I_g$ be the fraction of genes that are truly DE
- Category C (with m genes) contains a fraction $\pi_C = \frac{1}{m} \sum_{g \in C} I_g$ of DE genes
- Hence, we write $H_0 : \pi_C = \pi$ and $H_1 : \pi_C > \pi$
- We note that enrichment can then be defined by the quantity $\pi_C - \pi$.

Averaging v. Selection



- Each point is a single GO category
- 2761 GO categories plotted (each with $m \geq 10$)
- Significant correlation between Z_{ave} and Z_{sel}
- But many categories are outliers in only one method

Averaging v. Selection: Power comparison



- Consider one category with 20 genes
- $\pi = 0.20$
- red means low power
- Both methods increase in power as effect size increases or enrichment increases

Averaging v. Selection: Power comparison

Averaging

- A test based on $\bar{X}_{ave}$ has sampling distribution $N(\delta\pi_C, \frac{1}{m})$
- Hence, power of level-α test is $1 - \Phi(\tau_{ave})$, where

$$\tau_{ave} = z_\alpha - \sqrt{m}(\pi_C - \pi)\delta \tag{4}$$

Selection

- With a normal approximation, power for $\bar{X}_{sel}$ is $1 - \Phi(\tau_{sel})$, where

$$z_\alpha \frac{\sigma(\pi)}{\sigma(\pi_C)} - \sqrt{m}(\pi_C - \pi)[\Phi(k) - \Phi(k - \delta)]/\sigma(\pi_C) \tag{5}$$

- k is a function of π, δ and α^* and chosen to give a DFDR-controlled gene list at level α^*

Conclusions

- Random-set methods offer a more flexible approach than SAFE/GSEA and enable detection of distinct aspects of enrichment signal
- Within random-set methods, both selection and averaging strategies have regions of superiority that depend on enrichment, effect size δ, and number of genes in the GO category of interest

References

Bradley Efron, *Large-scale inference: empirical bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press, 2010.

Michael A Newton, Fernando A Quintana, Johan A Den Boon, Srikumar Sengupta, and Paul Ahlquist, *Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis*, The Annals of Applied Statistics (2007), 85–106.

Srikumar Sengupta, Johan A Den Boon, I-How Chen, Michael A Newton, David B Dahl, Meng Chen, Yu-Juen Cheng, William H Westra, Chien-Jen Chen, Allan Hildesheim, et al., *Genome-wide expression profiling reveals ebv-associated inhibition of mhc class i expression in nasopharyngeal carcinoma*, Cancer research **66** (2006), no. 16, 7999–8006.