Is a Set of Genes “Enriched”?

Bradley Efron

(Joint Work with Rob Tibshirani)

Hypothesis Testing in Microarray Experiments

- **Compare Genetic Activity** of thousands of genes under two different conditions:

  $Treatment_1$ (n_1, microarrays) *vs* $Treatment_2$ (n_2 microarrays)

- **Simultaneously Test** Null Hypotheses $H_0^{(i)}$: gene i has same activity under both conditions

Two-Sample test statistic "z_i" for gene i,

$$H_0^{(i)} : z_i \sim N(0,1)$$

Goal Identify a small number of "important" genes for further study

Trouble Not much information per gene, low power.

Gene-Set "Enrichment" (Subramanian et al., 2005)

• **Idea** Look at naturally defined *sets of genes*
rather than individual genes

Hope Biologically related gene-sets may reveal general z-value "enrichment" whether or not individual z_i's significant.

• **522 "Pathways"** $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{522}$, sizes $m = 2$ to 358,

• **GSEA Statistic** "gene-set enrichment analysis"
Kolmogorov-Simirnov distance between z_i's in $\mathcal{S}$
and all others [Mostly discuss getting p-value for a single $\mathcal{S}$.]

p53 Data
("transcription factor", NCI cancer cell lines)

- $N = 10,100$ *genes*, $n = 50$ microarrays,
$n_1 = 17$ "normal cell lines", $n_2 = 33$ "mutated"

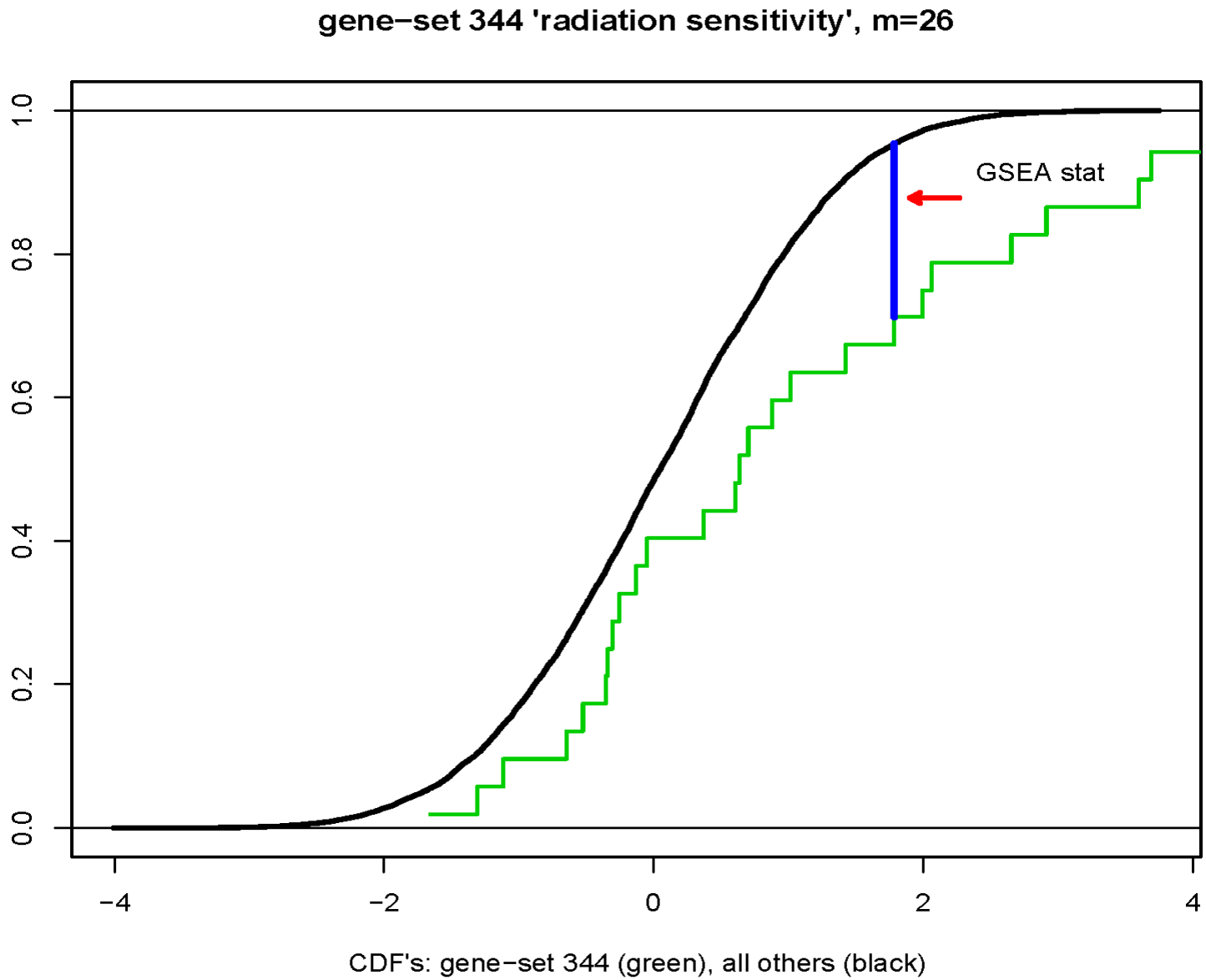
- *Expression Matrix* X is N by n

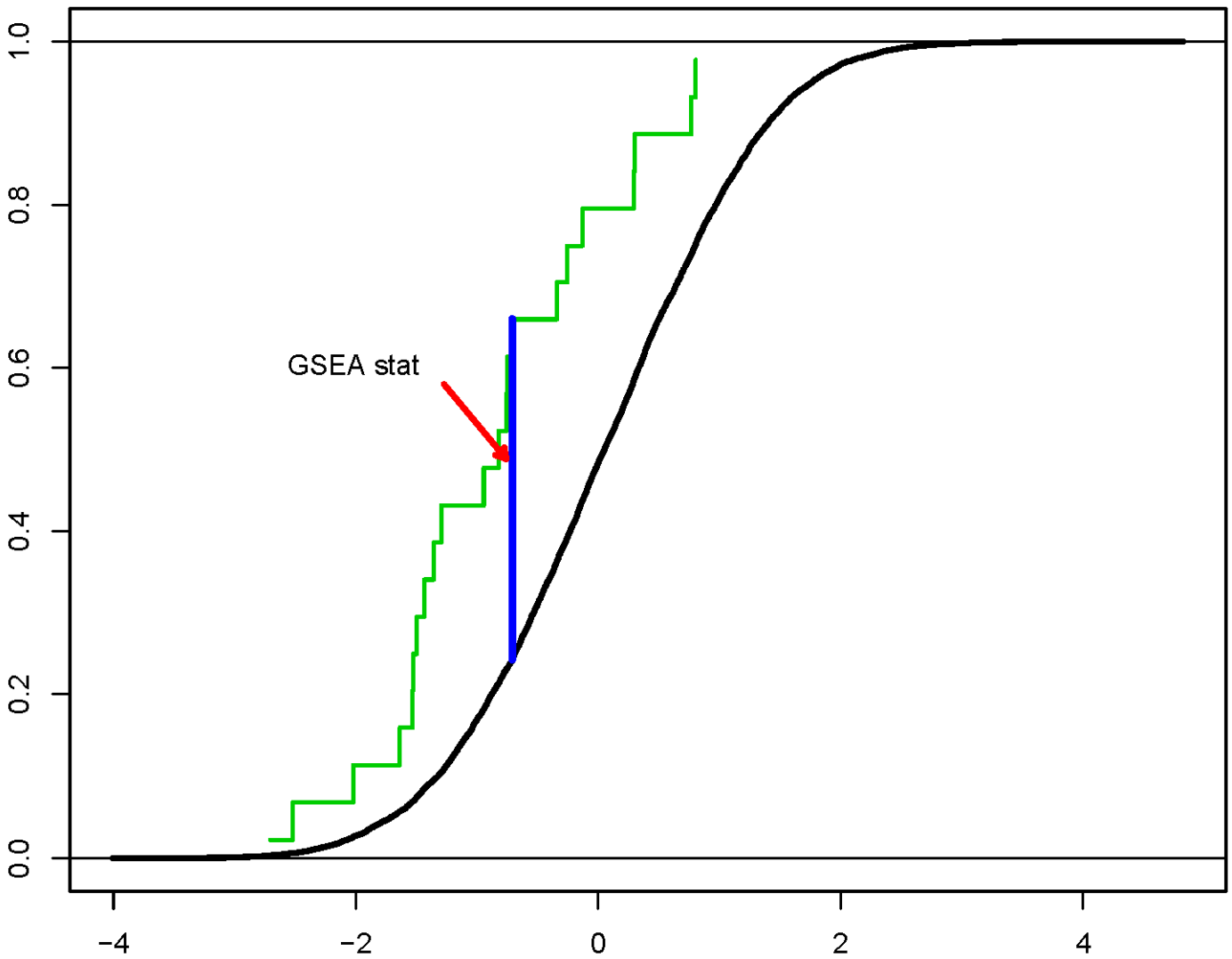
- *t-statistic* "t_i" comparing mutated vs normal for gene i

- *z-statistic* $\boxed{z_i = \Phi^{-1}(F_{n-2}(t_i))}$ $[\Phi, F_{n-2}$ normal, t_{n-2} cdfs$]$

so $z_i \sim N(0,1)$ under theoretical null.

- *Standard FDR analysis* shows only one non-null gene,
but GSEA gives 6 or 7 enriched gene-sets
(using permutations to get gene-set p-values).


gene−set 344 'radiation sensitivity', m=26
GSEA stat
1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
CDF's: gene−set 344 (green), all others (black)


gene−set 349 'rasPathway', m=22;
Signal Transduction Pathway
GSEA stat
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
CDF's: gene−set 349 (green), all others (black)

Other Enrichment Statistics (Efron & Tibshirani 2006)

- **Individual scores** $s_i = s(z_i)$

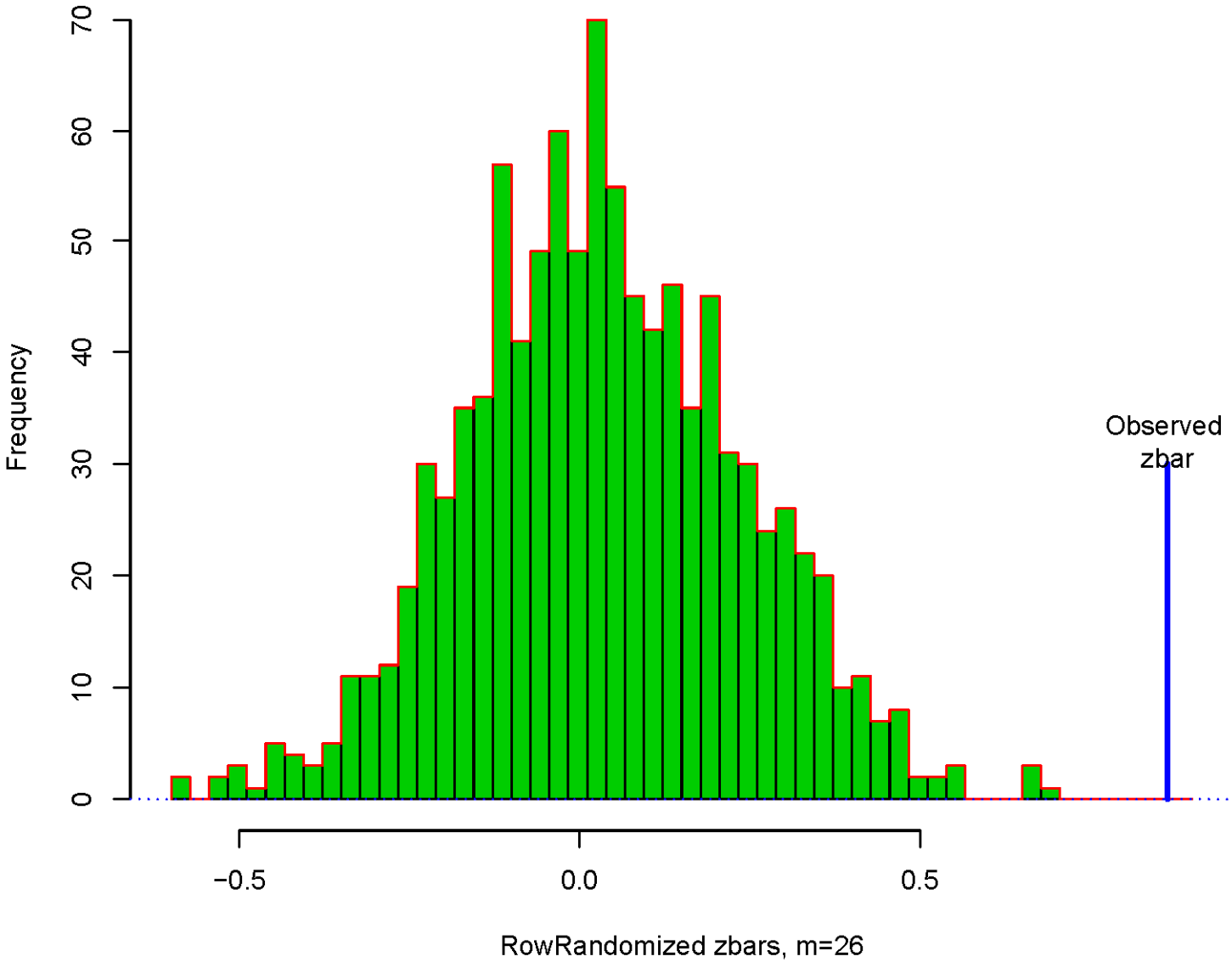
(1) $s_i = z_i$ (2) $s_i = |z_i|$ (3) $s_i = (\max(z_i, 0), -\min(z_i, 0))$

- **Enrichment Statistic** For gene-set $\mathcal{S}$, with m members,

$$S = \sum_{\mathcal{S}} s_i/m = \bar{s}_{\mathcal{S}}$$

- "*Limma*" (Smythe 2004) uses $s_i = z_i$. Compare $S = \bar{z}_{\mathcal{S}}$ to distribution of $S^\dagger = \bar{z}^\dagger_{\mathcal{S}}$, average of m random draws from all N z_i's:

"**Row Randomization**"
(Randomize rows of $N \times n$ matrix X, recompute $\bar{z}_{\mathcal{S}}$)


Compare 1000 RowRandomized zbars, m=26, with
observed zbar for gene−set 344
Frequency
0
10
20
30
40
50
60
70
Observed
zbar
−0.5
0.0
0.5
RowRandomized zbars, m=26

Efficient Enrichment Tests

- **Idea** Model choice of gene-set $\mathcal{S}$ *conditional* on $\mathbf{z}$, the set of all N z-values.

- *Poisson Selection Model* "Indicators" $I_i \overset{\text{ind}}{\sim}$ Poisson (v_i), $i = 1, 2, \ldots, N$,

$$v_i \propto e^{\beta s_i}, \; s_i = s(z_i) \; .$$

- *Suppose all* $I_i = 0$ *or* 1 · Selected genes-set

$$\mathcal{S} = \{i : I_i = 1\} \text{ has size } m = \sum_1^N I_i$$

- **Probability of Selection**

$$g_\beta(\mathcal{S}|m) = m! e^{m[\beta \bar{s}_\mathcal{S} - \psi(\beta)]}$$

$$\begin{cases} \bar{s}_\mathcal{S} = \sum_\mathcal{S} s_i/m \\ \psi(\beta) = \log \sum_1^N e^{\beta s_i} \end{cases}$$

Optimal Testing For Enrichment

- $H_0 : \beta = 0$ makes all gene-sets $\mathcal{S}$ of size m equally likely (i.e. Row Randomization)
- **Optimal Test Statistic** $S = \bar{s}_{\mathcal{S}}$
  - $\beta \neq 0$ means non-random selection.

(1) $s(z) = z, S = \bar{z}_{\mathcal{S}}$: good versus location alternatives

($\mathcal{S}$ has z_i's shifted left or right)

(2) $s(z) = |z|$, $S = \text{mean}_{\mathcal{S}}\{(z_i)\}$: good versus scale alternatives

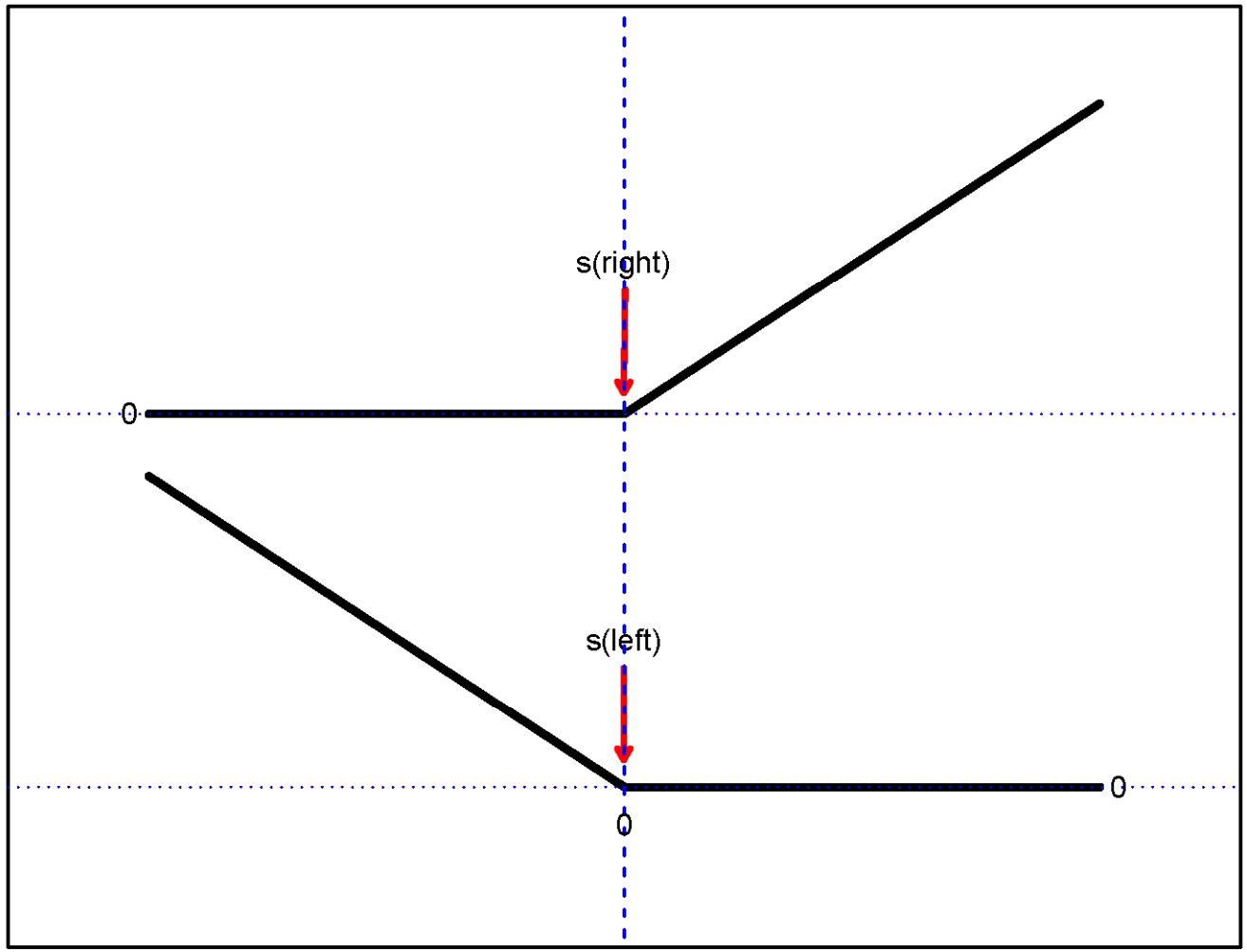
($\mathcal{S}$ has z_i's scaled up.)

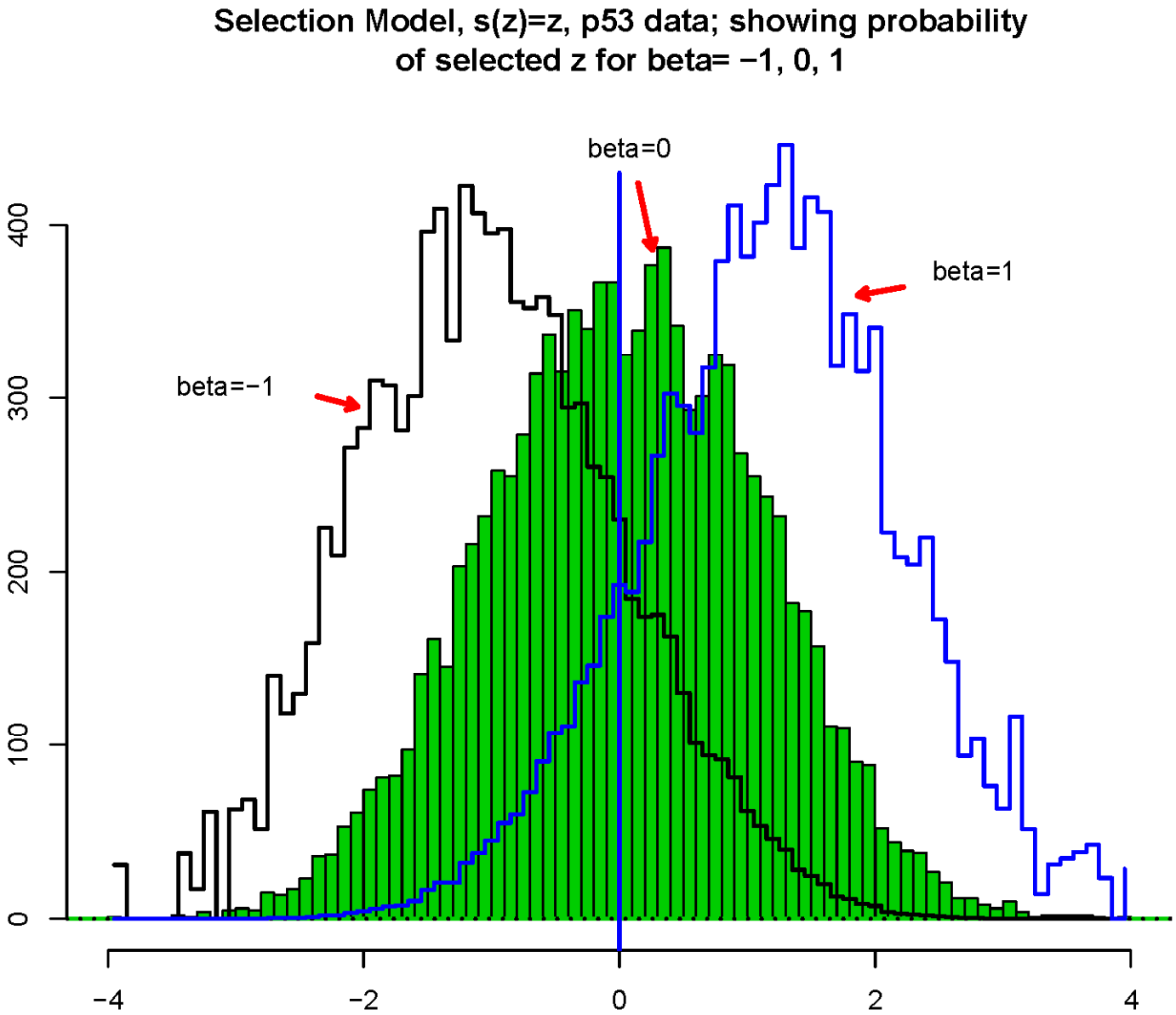
(3) $s(z) = (s^{\text{right}}(z), s^{\text{left}}(z)) : s^{\text{right}} = \max(z, 0),\ s^{\text{left}} = -\min(z, 0)$

Use $S = \max\{\bar{s}^{\text{right}}_{\mathcal{S}}, \bar{s}^{\text{left}}_{\mathcal{S}}\}$ to test both location, scale

"maxmean statistic"

The two maxmean s functions




Selection Model, s(z)=z, p53 data; showing probability
of selected z for beta= −1, 0, 1
beta=0
beta=1
beta=−1
400
300
200
100
0
−4
−2
0
2
4

Power Comparison (Simulation)

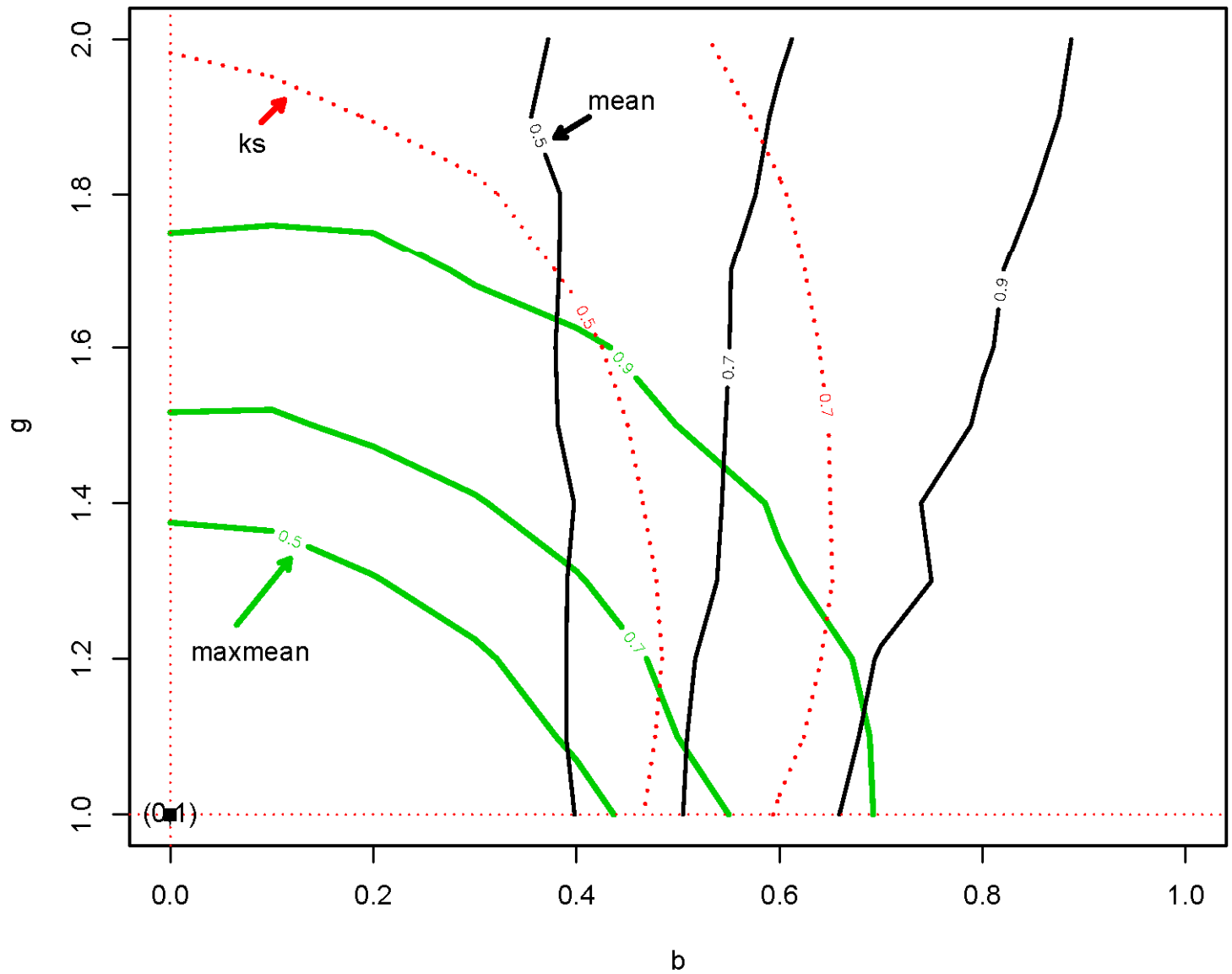
- *Simulate* gene-sets $\mathcal{S}$ with $m = 25$ members, z_i's i.i.d. normal
- Test $H_0 : z_i \sim N(0,1)$ vs $H_1 : z_i \sim N(b, g^2)$ using

  $S = \bar{z}_{\mathcal{S}}$, $S =$ maxmean, $S =$ Kolmongorov-Smirnov

- Next figure shows contours of (b, g) with power $= 0.5$

  (two-sided size $= 0.95$)

- Better Power $=$ Contour closer to (0, 1)


Power contours for Mean, MaxMean,
and KS Enrichment stats
ks
mean
maxmean
0.5
0.7
0.9
(0,1)
g
b
1.0
1.2
1.4
1.6
1.8
2.0
0.0
0.2
0.4
0.6
0.8
1.0

What About Permutations?

- GSEA compares observed S for $\mathcal{S}$ with permutation distribution:

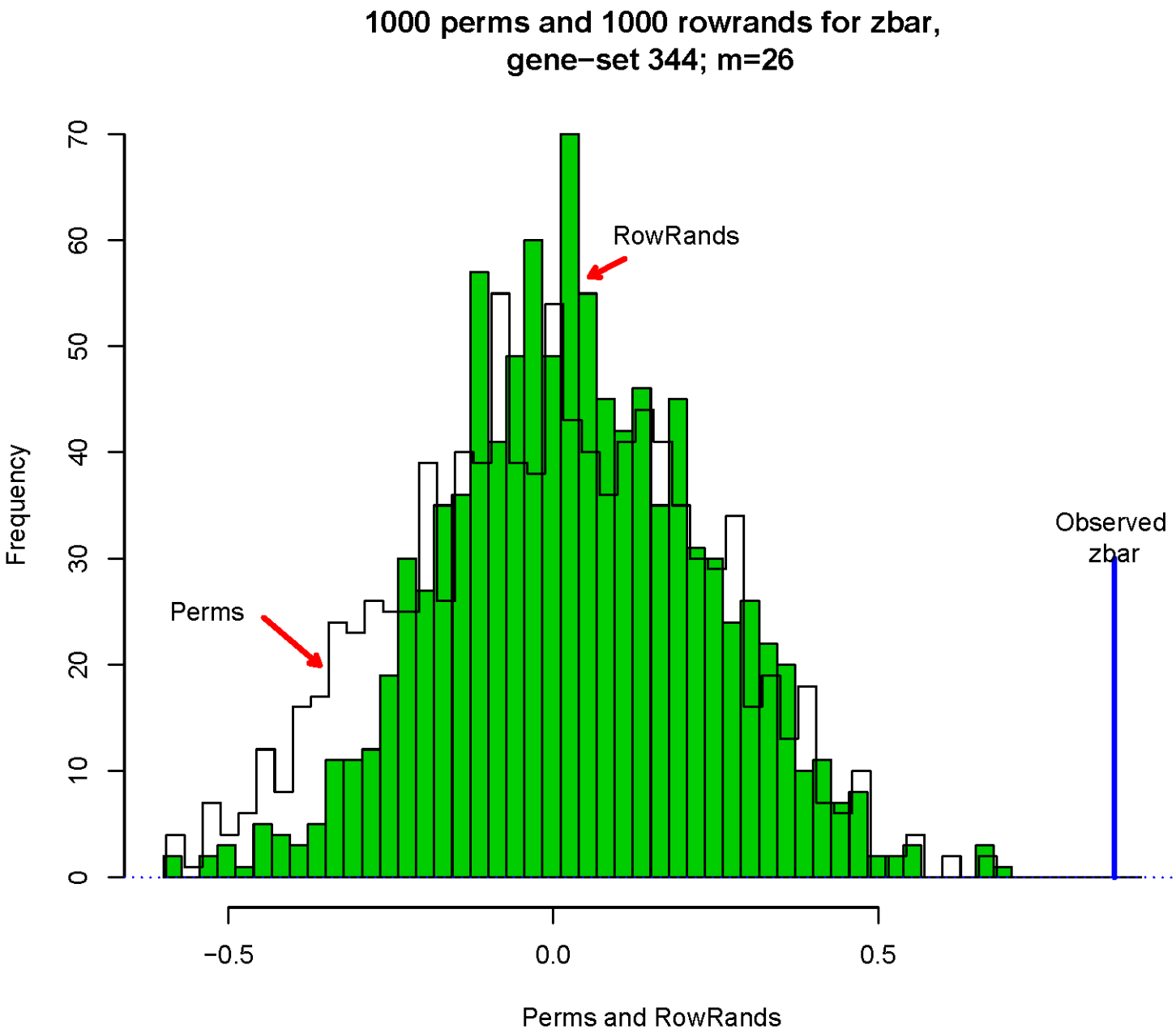
  (1) Randomize **Columns** $X_{N\times n}$ (2) Recompute $\mathbf{z}^*$
  (3) Recompute S^* for $\mathcal{S}$

- Row Randomization assumes genes in $\mathcal{S}$ are independent

- If genes in $\mathcal{S}$ positively dependent S will have greater variability

- Suppse $z_i \sim N(0,1)$ for m genes in $\mathcal{S}$, average correlation $\bar{\rho}$:

$$\text{var}(\bar{z}_{\mathcal{S}}) = \frac{1}{m}[1 + (m-1)\bar{\rho}]$$

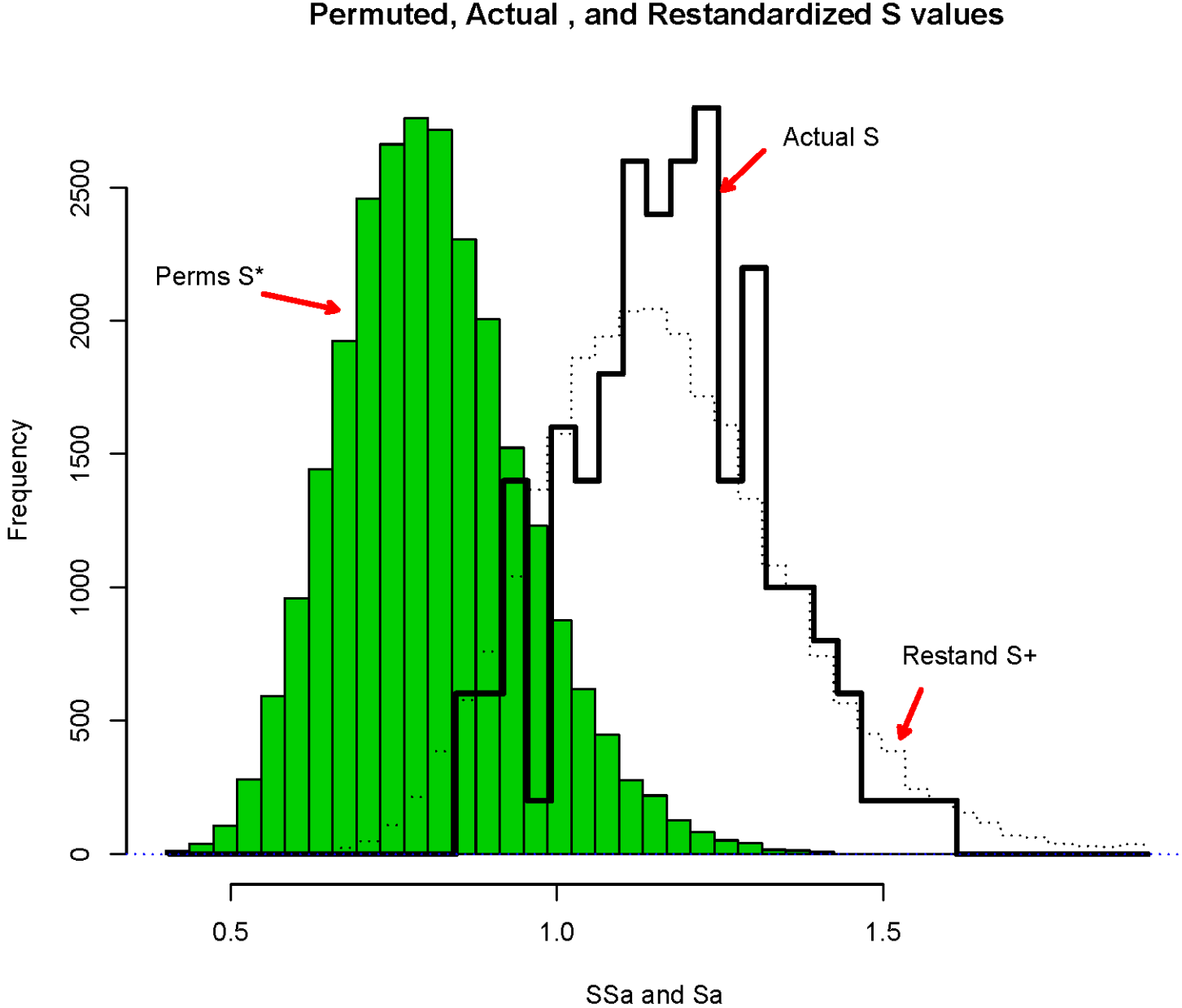

1000 perms and 1000 rowrands for zbar,
gene−set 344; m=26
Frequency
0
10
20
30
40
50
60
70
RowRands
Perms
Observed
zbar
−0.5
0.0
0.5
Perms and RowRands

Why Not Just Use Permutations?

BRCA Data (Efron 2004) · $N = 3226$ genes, $n = 15$ microarrays

$n_1 = 7$ "BRCA1" and $n_2 = 8$ "BRCA2" • X 3226 by 15

- *2-Sample t-stats* "t_i" give z-stats $z_i = \Phi^{-1}F_{13}(t_i)$ for $i = 1, 2, \ldots 3226$
- *Empirical distribution* of z_i's has $\boxed{(\widehat{\mu}, \widehat{\sigma}) = (-.03, 1.43)}$
- **Randomly Selected** $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{129}$, each with $m = 25$ genes
- *Enrichment Statistic* $S = \text{mean}_{\mathcal{S}}\{|z_i|\}$


Permuted, Actual , and Restandardized S values
Actual S
Perms S*
Restand S+
Frequency
0
500
1000
1500
2000
2500
0.5
1.0
1.5
SSa and Sa

Permutation Analysis of BRCA Gene-Sets

- *FDR analysis* of Perm p-values gives 113 Enriched gene-sets
  (But actually none are.)

- **Trouble** Perm distribution of 3226 z-values has (mean stdev)

$$(\mu^*, \sigma^*) = (.02, 1.01)$$

so Perm z_i^*'s much less dispersed than actual z_i's.

Permutations Preserve gene-wise correlations,
but lose ensemble z stats

Row Randomizations Lose correlations but preserve z stats.

Restandardization

- Let $(\widehat{\mu}_s, \widehat{\sigma}_s)$ be empirical (mean, sd) for all N values $s_i = s(z_i)$, and likewise

  (μ^*_s, σ^*_s) from preliminary perm calculations

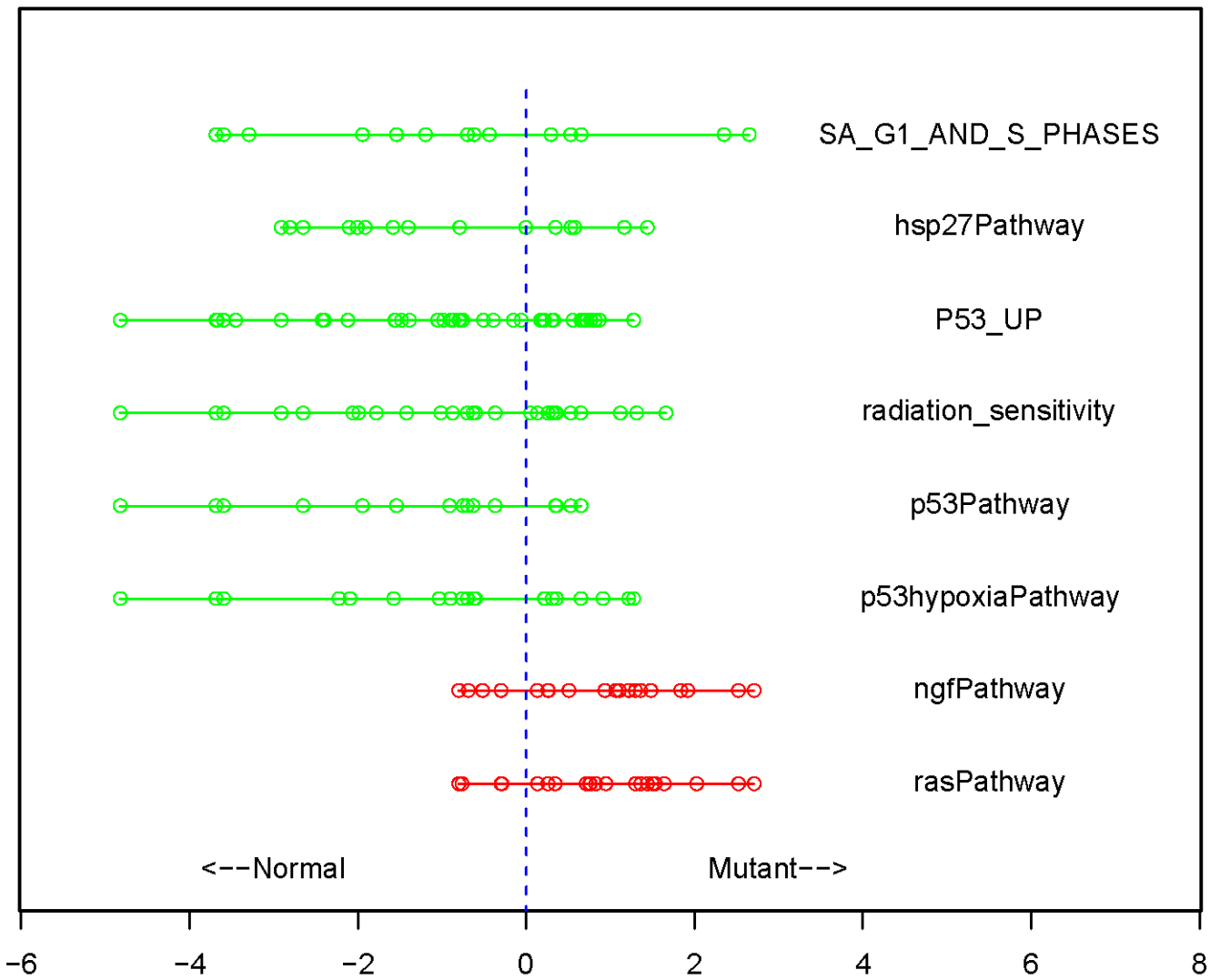
- *For given* $\mathcal{S}$, $S = \text{mean}_{\mathcal{S}}(s_i)$, get perm values $S^*(1), S^*(2), \ldots, S^*(B)$

- **Restandardized p-value**

$$p_{\mathcal{S}} = \#\left\{ \frac{S^* - \mu^*_s}{\sigma^*_s} > \frac{S - \widehat{\mu}_s}{\widehat{\sigma}_s} \right\} \Big/ B$$

- *Equivalently* $p_{\mathcal{S}} = \#\{S^{**} > S\}/B$ where $S^{**} = \widehat{\mu}_s + \frac{\sigma^*_s}{\widehat{\sigma}_s}(S^* - \mu^*_s)$

R Package "GSA"



Same as GSEA (which is approximately restandardized)

Two Different Null Hypotheses

- **Randomization Null** Members of $\mathcal{S}$ chosen randomly from all N genes,

  (Row Randomization, Poisson Selection Model)

- **Permutation Null** Columns of $X_{\mathcal{S}}$, the m by n matrix for $\mathcal{S}$, are i.i.d.

  (Permutation Distribution)

- **Restandardization** works well if
  - $\mathcal{S}$ selected randomly
  - The $N(0,1)$ null agrees with empirical distribution of z_i's
  - z_i's are uncorrelated

- But not a general solution!

Brain-Scan Study
(Armin Schwartzman et al. 2006)

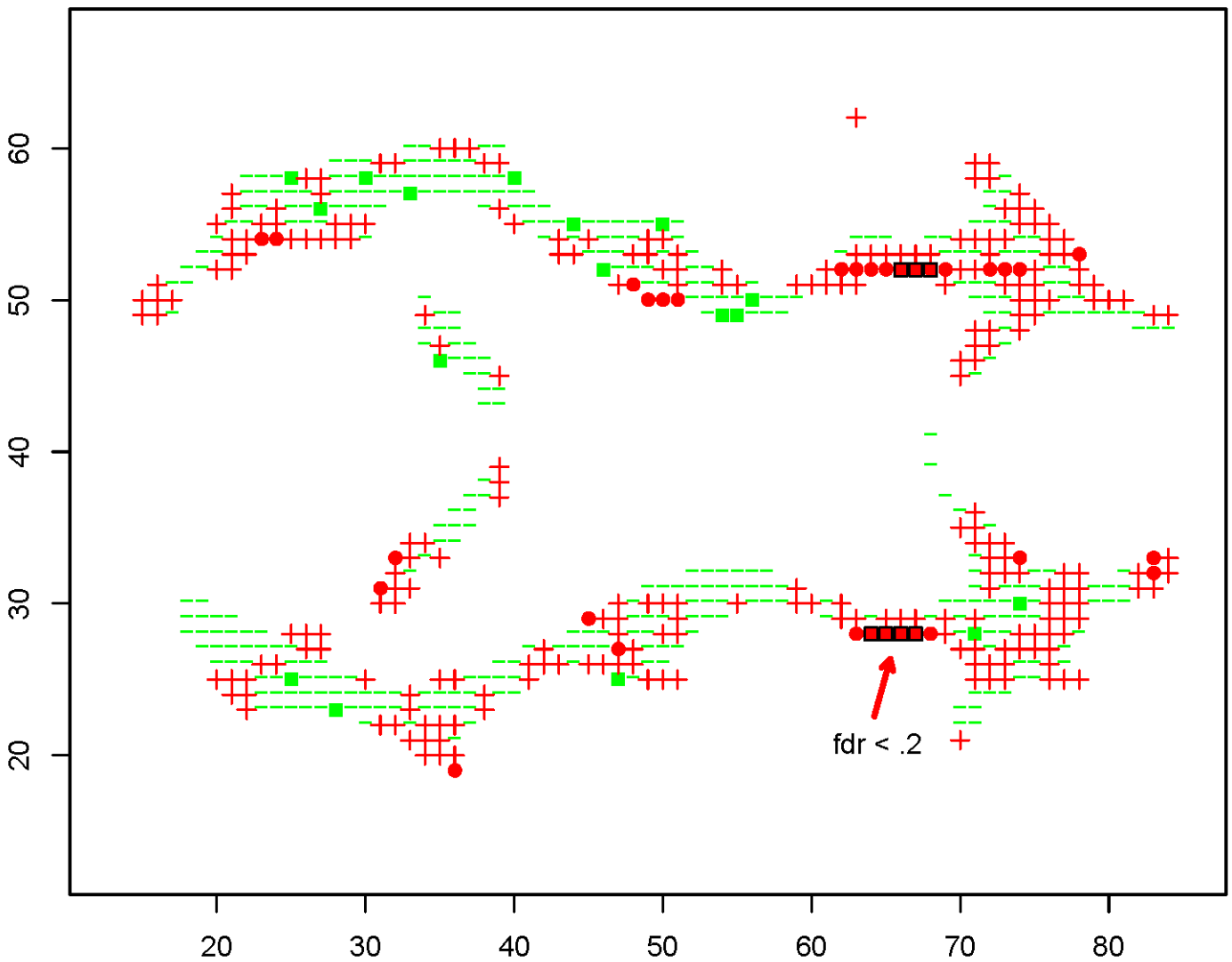
- *Diffusion Tensor Images* 12 children, $n_1 = 6$ normal, $n_2 = 6$ dyslexia

2-Sample t-stats give z_i's at $N = 15443$ voxels
  - Next figure shows one horizontal slice, 655 voxels.

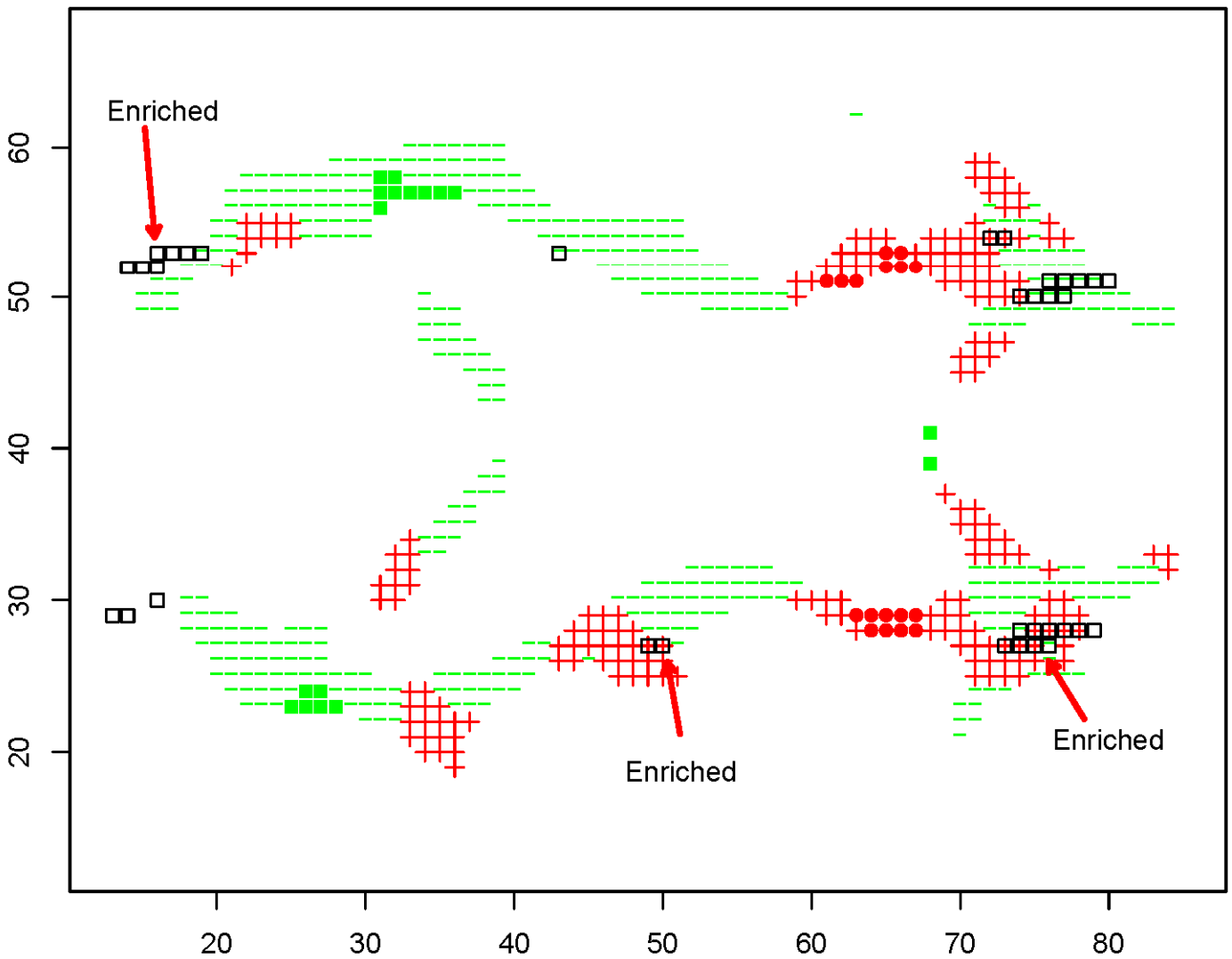
- **Gene-Sets** For voxel k of 655: $\mathcal{S} =$ voxels city block distance ≤ 2

  (average $m = 79$)

- *Limma statistic* is smoothed average $\bar{z}_{\mathcal{S}}$.



Schwartzman, Dougherty, and Taylor 2006


Now using zbar for city−block 2 gene−sets;
Enriched Areas indicated by Black Squares
Enriched
Enriched
Enriched
20
30
40
50
60
70
80
20
30
40
50
60

References

- Efron and Tibshirani "On testing the significance of a set of genes.
- Subramarian A., Tamaya, P., ...(....), Mesirov, J. (2005). "Gene set enrichment analysis ... " *PNAS* **102** 15545-15550.
- Smyth, G. (2004). Linear models and empirical Bayes methods ..." *Statistical Applications in Genetics and Molecular Biology* **3** (1).
- Schwartzman, A., Dougherty, R., Taylor, J. (2005). "Cross-subject comparison of principal diffusion direction maps", *Magn. Reson. Med.* To appear.
- Rahrenfuker, J., ..., Lengaver, T. (2004). "Calculating the statistical significance of changes in pathway activity from gene expression data", *Statistical Applications in Genetics and Molecular Biology* **3**, (16).
- Beisbarth, T., Speed, T. (2004). "GOstat: Find statistically over-represented Gene Ontologies within a group of genes", *Bioinformatics* **20**, 2468-65.
- Tian, Greenberg et al., (2005). *PNAS* **102**, 13544-49.