第6-2章: Class Comparison

1. Statistical test
2. Parametric empirical Bayesian method
3. Gene set enrichment analysis (GSEA)

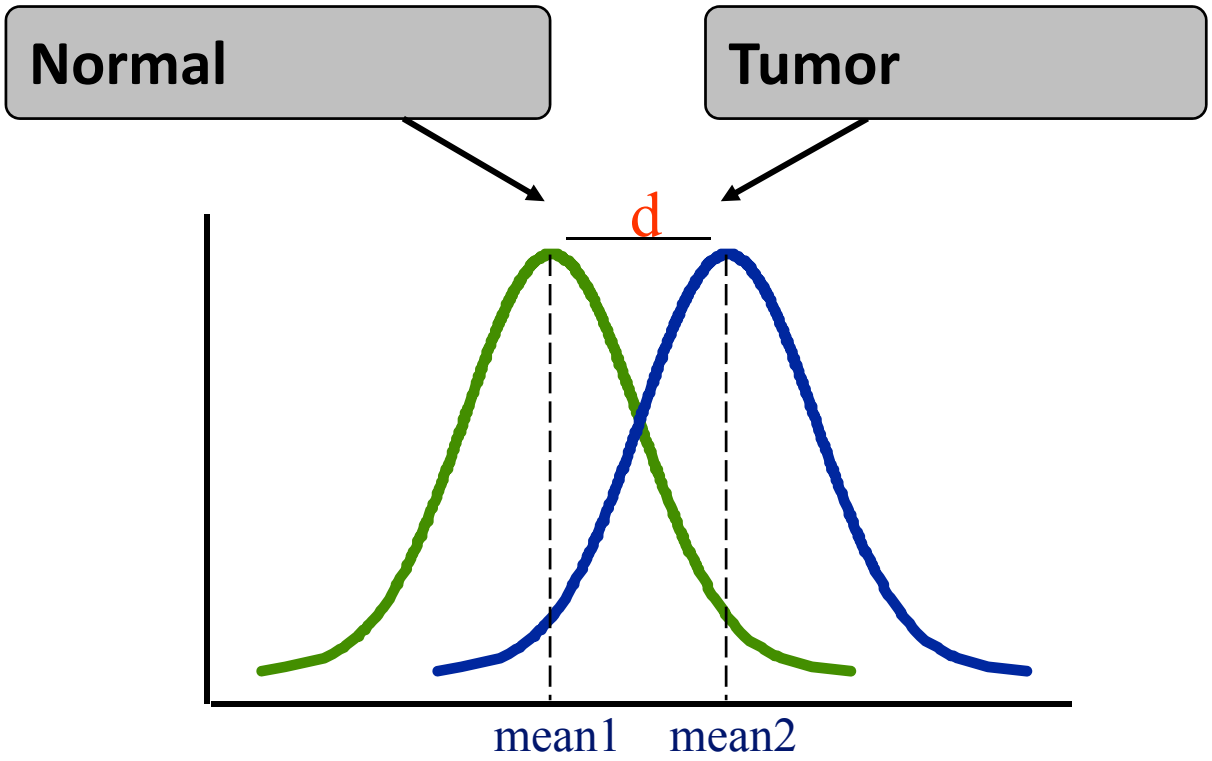
Class comparison

- What genes are up regulated between control and test or multiple test conditions
  - Normal vs tumor
  - Treated vs untreated
- Fold change
  - Not sufficient, need statistics
- Statistics
  - t test, non-parametric, fdr
- Depends on underlying assumptions about data

Class Comparison

- What genes are up regulated between control and test or multiple test conditions
- Many analysis methods
  - May produce different results
  - Different underlying statistics and methods
    - t test
    - SAM
    - Non parametric (relative entropy)
    - Emperical bayesian
- Depends on underlying assumptions about data

Hypothesis Testing



Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypotheses $H_1 : \mu_1 \neq \mu_2$

Type I and Type II Error

	Retain Null	Reject Null
H_0	$\surd$	type I error
H_1	type II error	$\surd$

Two-Sample *t*-Statistic

- Student's *t*-statistic (equal variance)

$$T = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

- Student's *t*-statistic (unequal variance)

$$T = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Multiple Test Problem

- Perform a test for each gene to determine the statistical significance of differential expression for that gene.

- Problem: When many hypotheses are tested, the probability of a type I error (false positive) increases sharply with the number of hypotheses.

Example

- Suppose we measure the expression of 10,000 genes in a microarray experiment.
- p <= 0.05 says that 95% confidence means are different; therefore 5% due to chance
- 500 genes are picked up by chance
- Suppose t tests selects 1000 genes at a p of 0.05
- 500/1000 ;Approximately 50% of the genes will be false, very high false discovery rate; need more confidence

Corrections for Multiple Comparisons

- Involve corrections to the p-value so that the actual p-value is higher
- Bonferroni correction
- Benjamin-Hochberg procedure
- Significance Analysis of Microarrays
  - Tusher et al. at Stanford

The Bonferroni Method

- Controls the family wise error rate (FWER) FWER is the probability that at least one false positive error will be made.

- But this method is very conservative, as it tries to make it unlikely that even one false rejection is made.

False Discovery Rate (FDR)

- The FDR is essentially the expectation of the proportion of false positives among the identified differentially expressed genes.

$$\mathrm{FDR} \approx \frac{\#(\text{False Positives})}{\#(\text{Rejected Hypotheses})}$$

False Discovery Rate

	Accept Null	Reject Null	Total
Null True	N_{00}	N_{01}	N_0
Non-True	N_{10}	N_{11}	N_1
Total	$N - N_r$	N_r	N

$$FDR \approx \frac{N_{01}}{N_r}$$

Benjamini-Hochberg (BH) Procedure

Controls the FDR at level α when the *P*-values following the null distribution are independent and uniformly distributed.

(1) Let $p_{(1)} \leq \cdots \leq p_{(N)}$ be the observed *P*-values.

(2) Calculate $\hat{k} = \underset{1 \leq k \leq N}{\text{Argmax}} \{p_{(k)} \leq \frac{k}{N}\alpha\}$

(3) If $\hat{k}$ exists then reject null hypotheses corresponding to

$p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

Example: Bonferroni and BH Tests

Suppose that 10 independent hypothesis tests are carried out leading to the following ordered *P*-values:

$$0.00017, 0.00448, 0.00671, 0.00907, 0.01220$$
$$0.33626, 0.39341, 0.53882, 0.58125, 0.98617$$

(a) With α = 0.05, the Bonferroni test rejects any hypothesis whose *P*-value is less than α / 10 = 0.005.

Thus only the first two hypotheses are rejected.

(b) For the BH test, we find the largest *k* such that $P_{(k)} < k\alpha / m$.

Here *k* = 5, thus we reject the first five hypotheses.

Using just the *B* permutations of the class labels for the gene-specific statistic T_j, the *P*-value for $T_j = t_j$ is assessed as:

$$p_j = \frac{\#\{b : t_{0j}^{(b)} \geq t_j\}}{B}$$

where $t_{0j}^{(b)}$ is the null version of t_j after the *b*th permutation of the class labels.

Permutation Test

- The null distribution of statistic
- Generally one can permute the sample label, and calculate the statistic. And repeat such a procedure B times
- If we perform B permutations, then the P-value will be estimated with a resolution of 1/B.

Hierarchical Bayesian Model
Parametric empirical Bayes methods

Ref： C. M. Kendziorski, M. A. Newton, H. Lan and M. N. Gould. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statist. Med. 22:3899-3914, 2003.

Hierarchical Bayesian Model

- Bayesian method treat genes as arising from some population, which allow a level of information sharing among genes. So it significantly reduce the dimensionality
- The hierarchical model can captures differences among genes in their average expression level, differential expression for a given genes among several cell types, and measurement fluctuations.

Hierarchical Bayesian Model

- Data: Replicate expression profiles in multiple conditions.
- Objective: Find the differentially expressed genes

- Method: posterior odds ratio of differential expression

General Mixture Model

- Suppose in the general case that m + 1 distinct patterns of expression are possible for a data vector $dg = (d_{g,1}, \cdots, d_{g,N})$ measuring a gene g in N conditions.
- For any pattern k, N experiment conditions are partitioned into r(k) mutually exclusive and exhaustive subsets $S_{i,k}, i = 1, \cdots, r(k)$
- Any measurements contained in a subset $S_{i,k}$ share a common latent mean level of expression

General Mixture Model

- Null hypothesis: all data share the same mean expression level, with distribution $f_0(d_g)$
- Alternative model: for pattern k, its distribution $f_k(d_g)$
- Distribution of data is given by a mixture model

$$\sum_{k=0}^{m} p_k f_k(d_g)$$

General Mixture Model

- Posterior probability of expression pattern k

$$Pr(k|d_g) \propto p_k f_k(d_g)$$

- Posterior odds

$$\frac{p_k}{1-p_k} \times \frac{f_k(d_g)}{1-f_k(d_g)}$$

Hierarchical Bayesian Model

- For replicates sharing a common latent mean expression level arise independently and identically from $f_{obs}(\cdot|\mu_g)$
- The latent mean can fluctuate among genes with a genome-wide distribution $\pi(\mu_g)$
- By integrating out the latent variable, we have

$$f(d_{g,s_{i,k}}) = \int \prod_{s \in S_{i,k}} f_{obs}(d_{g,s}|\mu_g)\pi(\mu_g)d\mu_g$$

Gamma-Gamma Model

- Gamma distribution with shape parameter a and scale parameter

$$f_{obs}(z|\mu_g) = \frac{\lambda_g^{\alpha} z^{\alpha-1} \exp(-\lambda_g z)}{\Gamma(\alpha)}$$

$$\lambda_g = \frac{\alpha}{\mu_g}$$

- Mean: $\mu_g = \alpha/\lambda_g$
- Variance: $1/\sqrt{\alpha}$

Gamma-Gamma Model

- Match to the Gamma distribution, we take inverse-Gamma distribution. More precisely, fixing α, the quantity $\lambda_g = \alpha/\mu_g$ has a Gamma distribution with shape parameter α_0 and scale parameter v.

Gamma-Gamma Model

- Then the distribution is

$$f(z_1, \cdots, z_n) = K \frac{(\prod_{i=1}^{n} z_i)^{\alpha-1}}{(v + \sum_{i=1}^{n} z_i)^{n\alpha+\alpha_0}}$$

$$K = \frac{v^{\alpha_0}\Gamma(n\alpha + \alpha_0)}{\Gamma(\alpha)^n \Gamma(\alpha_0)}$$

Inference

- Posterior odds for two groups

$$\text{odds}_g = \frac{p}{1-p} K' \frac{\left(\sum_{i=1}^{n_1} x_{g,i} + \sum_{i=1}^{n_2} y_{g,i} + v\right)^{N\alpha+\alpha_0}}{\left(\sum_{i=1}^{n_1} x_{g,i} + v\right)^{n_1\alpha+\alpha_0} \left(\sum_{i=1}^{n_2} y_{g,i} + v\right)^{n2\alpha+\alpha_0}}$$

$$K' = \frac{v_0^{\alpha}\Gamma(n_1\alpha+\alpha_0)\Gamma(n_2\alpha+\alpha_0)}{\Gamma(\alpha_0)\Gamma(N\alpha+\alpha_0)}$$

Empirical Bayesian

- The “empirical” means the small number of unknown parameters are estimated from the data
- Three parameters $\theta = (\alpha, \alpha_0, v)$ are estimated by a S-Plus program nlminb.

Gamma Distribution on Wiki

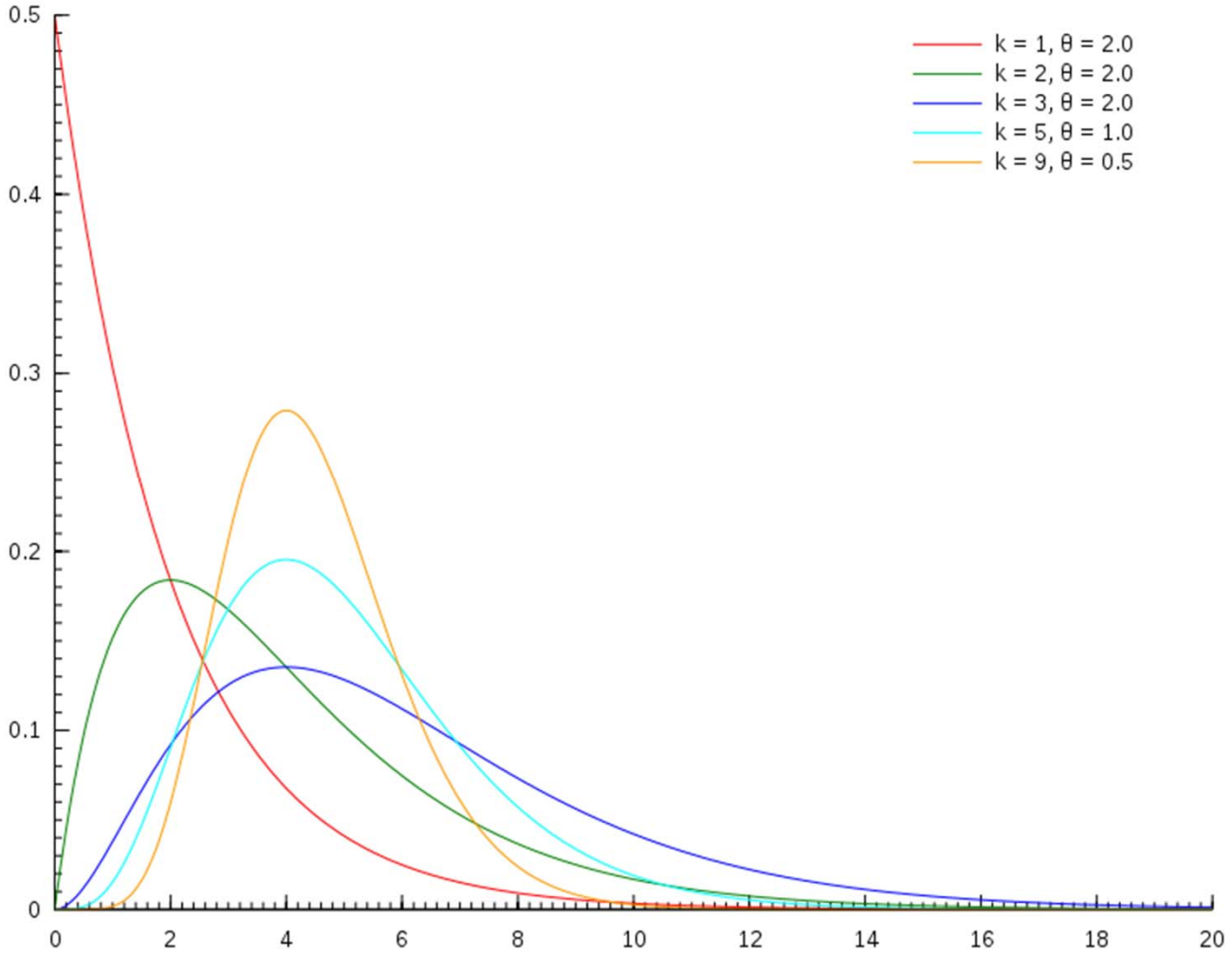
- In life test, the waiting time until death is generally model as Gamma distribution
- Density function (shape α , scale β)

$$\mathrm{Gamma}(x;\alpha,\beta) = \frac{x^{\alpha-1}\beta^{\alpha}e^{-x\beta}}{\Gamma(\alpha)}$$

- Expectation: $E[\ln(X)] = \psi(\alpha) - \ln(\beta)$
- Gamma distribution is the conjugate prior for many distributions

http://en.wikipedia.org/wiki/Gamma_distribution

Gamma Distribution on Wiki



http://en.wikipedia.org/wiki/Gamma_distribution

Inverse-Gamma Distribution on Wiki

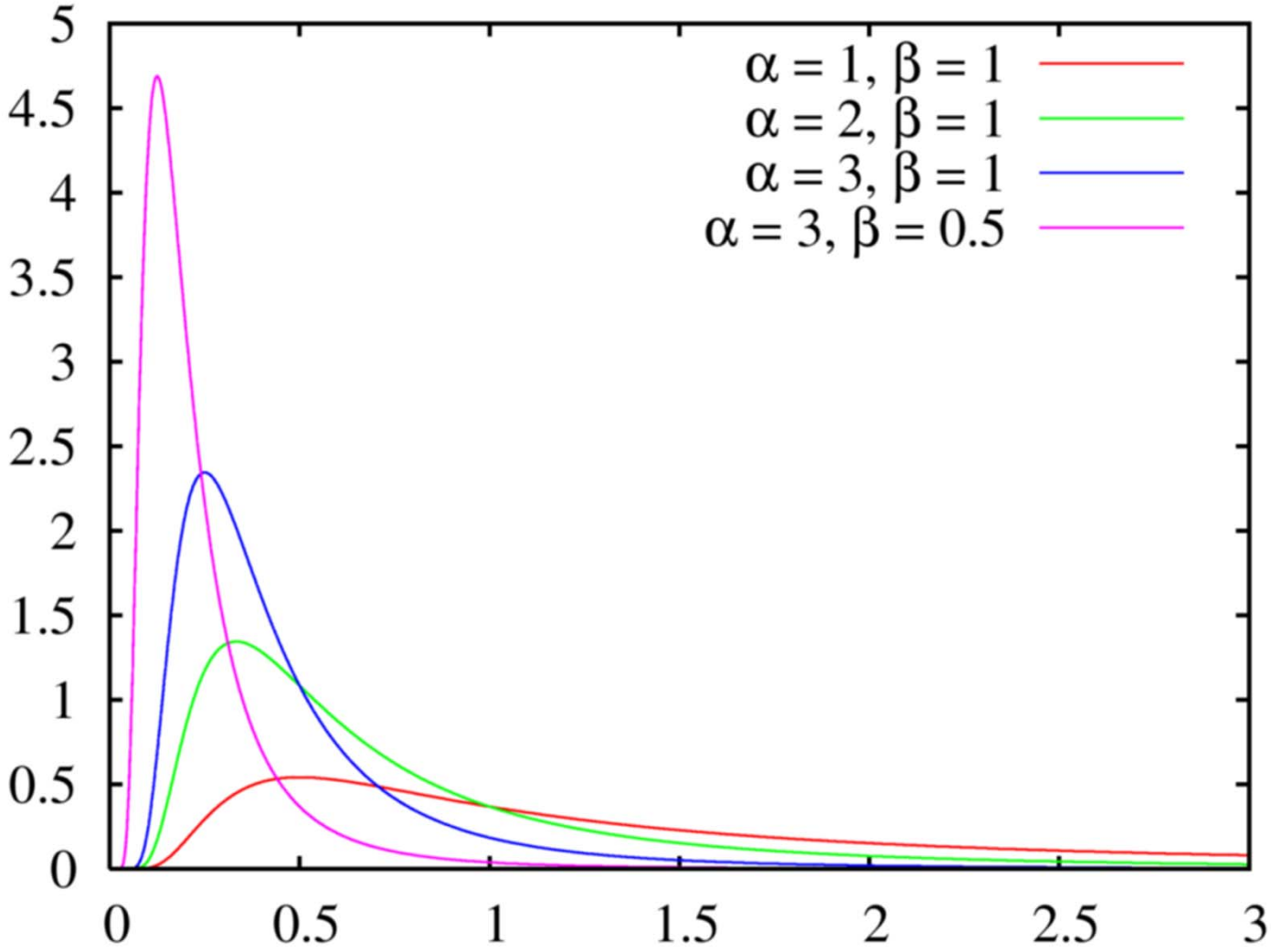
- Density function

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{-\alpha-1} \exp(-\frac{\beta}{x})}{\Gamma(\alpha)}$$

- E(1/X)=a/b
- If X~Inv-Gamma(a,b), kX~Inv-Gamma(a,kb)
- If X~Gamma(a,b), 1/X~Inv-Gamma(a,b)
- If X~Gamma(k,t), 1/X~Inv-Gamma(k,1/t)

http://en.wikipedia.org/wiki/Inverse-gamma_distribution

Inverse-Gamma Distribution on Wiki



http://en.wikipedia.org/wiki/Inverse-gamma_distribution

Gene Set Enrichment Analysis (GSEA)

References:

1. Subramanian et al. PNAS 102:15546, 2005.
2. Tian et al. PNAS 102:13544, 2005.
3. Mootha et al. Nature Genetics 2003

Motivation

- Interpreting the results to gain insights into biological mechanisms remains a major challenge;
- For a typical study (e.g., experimental condition vs. control, disease state vs. normal, tumor type A vs. tumor type B), a standard approach has been to produce a list of differentially expressed genes (DEGs);

Challenges in Interpreting Gene Microarray Data

- May obtain a long list of statistically significant genes without any obvious unifying biological theme;
- Even with DEG list(s) of up and/or down-regulated genes, still need to accurately extract valid biological inferences. Cutoff for inclusion in DEG lists is somewhat arbitrary. Must address multiple hypothesis testing.

An Existing Way to Study Enrichment of Gene Categories

- Statistical procedures such as ***Fisher's exact test*** based on the hypergeometric distribution are used to test if members of a list of differentially expressed genes are overrepresented in given GO categories or in predefined gene sets compared with the distribution of the whole set of genes represented on the chip.
- Tools developed along this line include:
  - GOMINER;
  - GENMAPP;
  - ONTO-TOOLs;
  - CHIPINFO;
  - GOSTAT.

Limitation of Above Methods

- No further use made of information contained in expression values for the non-DEG list genes
- The level of differential expression of the genes in the significant gene list is not taken into consideration.
- The correlation structure of the expression data is not considered at all.

Introduction of GSEA

- First explored in Mootha's *Nature Genetics (03)* paper, fully formulated in PNAS(05) paper.
- *GSEA:* evaluate microarray data at the level of gene sets, which is defined based on prior knowledge (such as gene sets from GO categories or pathways from KEGG).

Overview of GSEA

- Given a prior defined gene set S , *GSEA* is to determine whether members of S are randomly distributed throughout the list, or primarily found at the top or bottom in the list.
- Step of *GSEA*:
  - Calculation of an enrichment score (*ES*).
  - Estimation of significance level of *ES*.
  - Adjustment for *MHT*.

Calculation of ES

- Notation: D is the expression dataset with N genes and k samples; C is a phenotype or profile of interest; N_H is gene number of S,
- Rank order N genes to form $L=\{ g_1, \cdots, g_N\}$ according their correlation $r(g_j)=r_j$.
- Define:

$$P_{hit}(S,i) = \sum_{g_j \in S, j \le i} \frac{|r_j|^p}{N_R}, \; N_R = \sum_{g_j \in S} |r_j|^p,$$

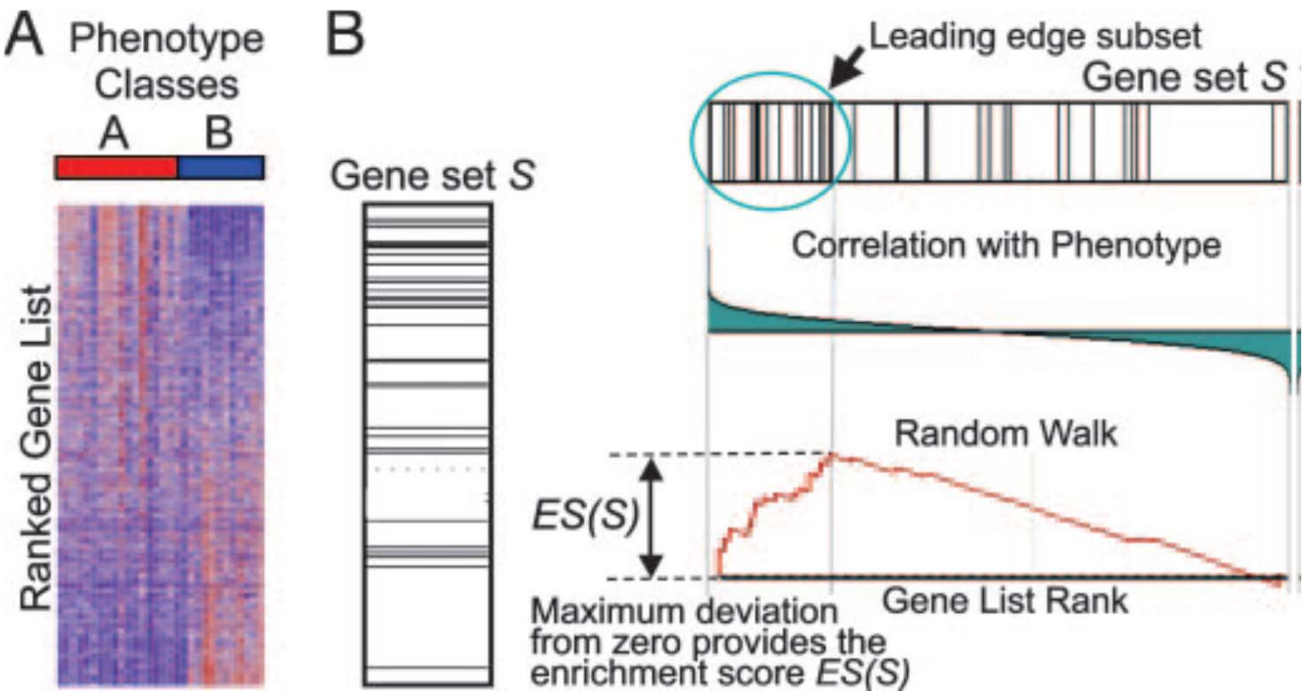
$$P_{miss}(S,i) = \sum_{g_j \notin S, j \le i} \frac{1}{N - N_H}$$

Where p is a constant to control the weight of ranks.

Calculation of ES

- The *ES* is the maximum deviation from zero of P_{hit}-P_{miss}.
- For a randomly distributed S, ES(S) will be small, but if it is concentrated at the top or bottom of the list, the score will be high.
- When p=0, this reduces to the standard *Kolmogorov-Smirnov statistic*.
  - As P_{hit} is the empirical distribution for genes in S, while P_{miss} is the one for genes outside S.

GSEA Overview



Effect of Weight p.

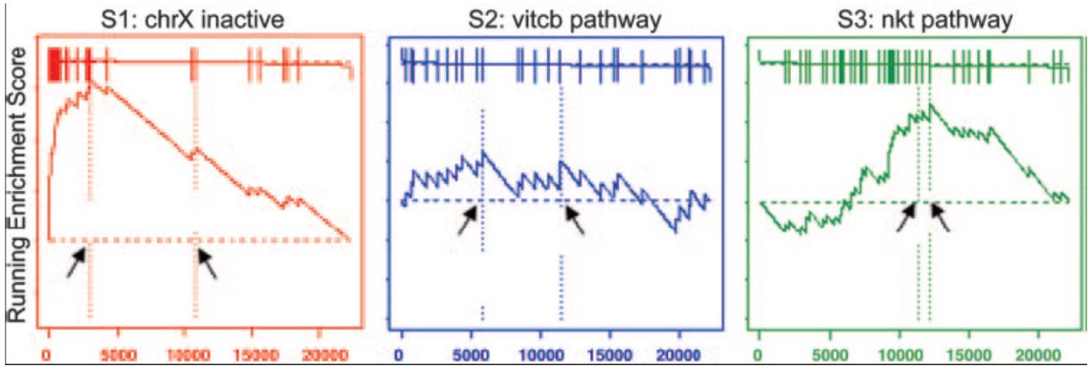


Table 1. *P* value comparison of gene sets by using original and new methods

Gene set	Original method nominal *P* value	New method nominal *P* value
S1: chrX inactive	0.007	<0.001
S2: vitcb pathway	0.51	0.38
S3: nkt pathway	0.023	0.54

Estimating Significance

- Randomly assign the original phenotype labels to samples, reorder genes, re-compute ES(S).
- Repeat for 1000 permutations, and create a histogram of the corresponding ES_{NULL};
- Estimate nominal p-value for S from ES_{NULL} and observed ES(S).

Multiple Hypothesis Test (MHT)

- Determine ES(S) for each gene set in database
- For each S and 1000 fixed permutations π of the phenotype labels re-order the genes in L and determine ES(S,π).

$$\begin{pmatrix} ES(S_1,\pi_1) & ES(S_1,\pi_2) & \cdots & ES(S_1,\pi_{1000}) \\ ES(S_2,\pi_1) & ES(S_2,\pi_2) & \cdots & ES(S_2,\pi_{1000}) \\ \cdots & \cdots & \cdots & \cdots \\ ES(S_K,\pi_1) & ES(S_K,\pi_2) & \cdots & ES(S_K,\pi_{1000}) \end{pmatrix}$$

Multiple Hypothesis Test (MHT)

- Normalize the ES(S,π) and observed ES(S), yields the normalized scores NES(S,π), NES(S).
- Compute FDR, for a given NES(S)=NES$^{*} \geq 0$.

$$\frac{\#\text{ of all } (S,\pi) \text{ with } NES(s,\pi) \geq 0 \text{ whose } NES(S,\pi) \geq NES*}{\#\text{ of observed } S \text{ with } NES(s,\pi) \geq 0 \text{ whose } NES(S,\pi) \geq NES*}$$

Results on p53 status in NCI-60 cell lines

- Enriched in p53 mutant
  - Ras signaling pathway 0.171
- Enriched in p53 wild type
  - Hypoxia and p53 in the cardiovascular system 0.001
  - Stress induction of HSP regulation 0.001
  - p53 signaling pathway 0.001
  - p53 up-regulated genes 0.013
  - Radiation sensitivity genes 0.078