Introduction to pathway analysis based on high-throughput genomic data

Shih Sheng Jiang

National Institute of Cancer Research

NHRI

台灣生物資訊核心設施
Taiwan Bioinformatics Institute Core Facility

I-Shou Chang[1], Ping-Chiang Lyu[2]
Chao A. Hsiung[1], Hsien-Da Huang[3], H. Sunny Sun[4]

**1 National Health Research Institutes, 2 National Tsing Hua University,
3 National Chiao Tung University, 4 National Cheng Kung University**

TBI Core Facility 服務內容

- 線上服務
- 客製化服務
- 諮詢服務

客製化服務項目

- **1.**微陣列數據分析

- **2.**新一代定序（**NGS**）資料分析

- **3.**基因體註解及功能性分析

- **4.**整合型高通量基因體數據生物標記選取服務

- **5.**蛋白質-化合物與藥物設計整合型分析服務

- **6.**基因體比較分析

- **7.**遺傳研究分析

- **8.**病毒/微生物之生物資訊學客製化分析

- **9.**資料庫設計開發

Based on

- Mootha et al. (2003) PGC-1α responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 34, 3, 267-273

- Subramanian et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, 102, 15545-15550

- Tamayo et al. (2012) Gene set enrichment analysis made right.

Mootha et al. (2003)

- In a follow-up study, 17 NGT(normal glucose tolerance), 8 IGT(impaired glucose tolerance, 18 DM2. Age-matched.

- Diagnosed by hyperinsulinemic clamp. At time of diagnosis, thus before treatment, skeletal muscle biopsy samples were taken.

- Microarray expression experiments.

- No single gene had a significant difference in expression between the diagnostic categories.

- Given a gene set (pathway), they ask if it is associated with the disease status.

Basic ideas

- Consider only NGT and DM2 to simplify the matter.

- Order the genes according to their differences in expression.

- A given gene set whose members contain more low ranking or high ranking genes than expected by chance.

- Collected databases of gene sets, including pathways.

- Developed statistical methods.

The findings

- 149 gene sets were prepared without knowledge of the expression data.

- The gene set getting highest score is OXPHOS (oxidative phosphorylation)

- The next four highest scoring gene sets overlap with the top one. This overlap explains the enrichments greatly.

- OXPHOS has 106 genes. 94 of them showed a decrease of about 20% in DM2 subjects.

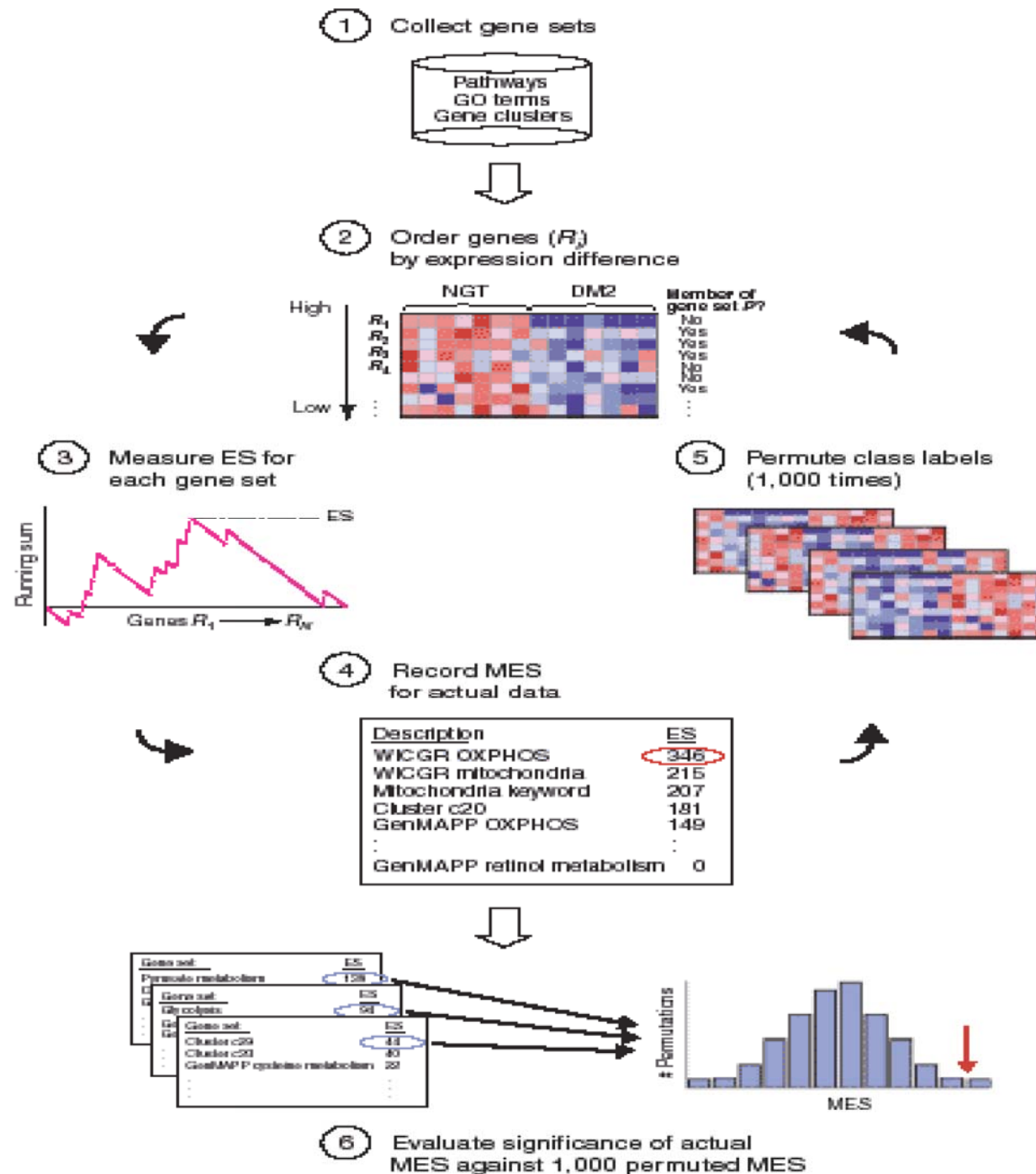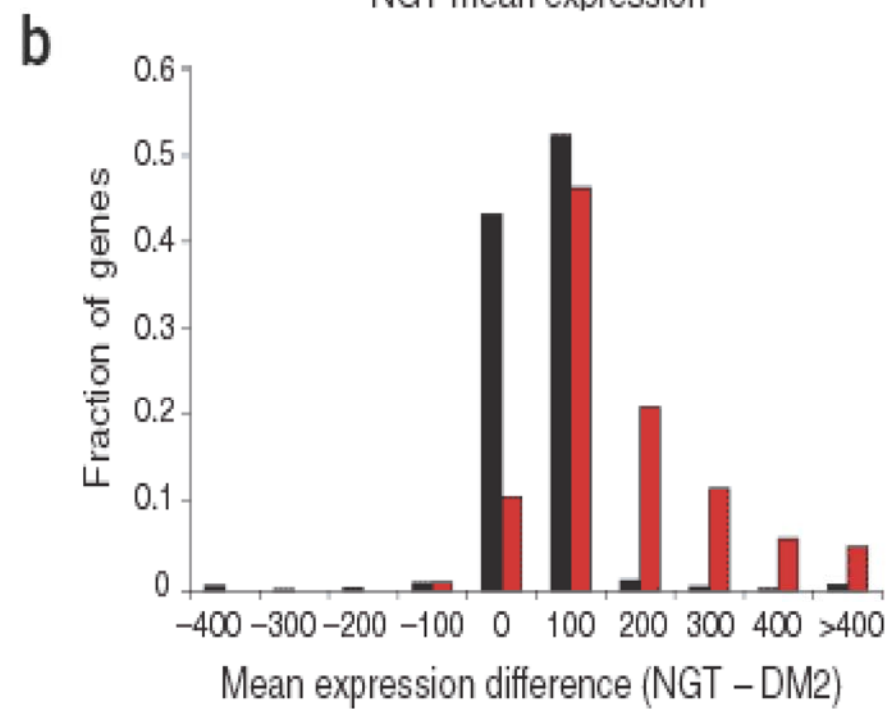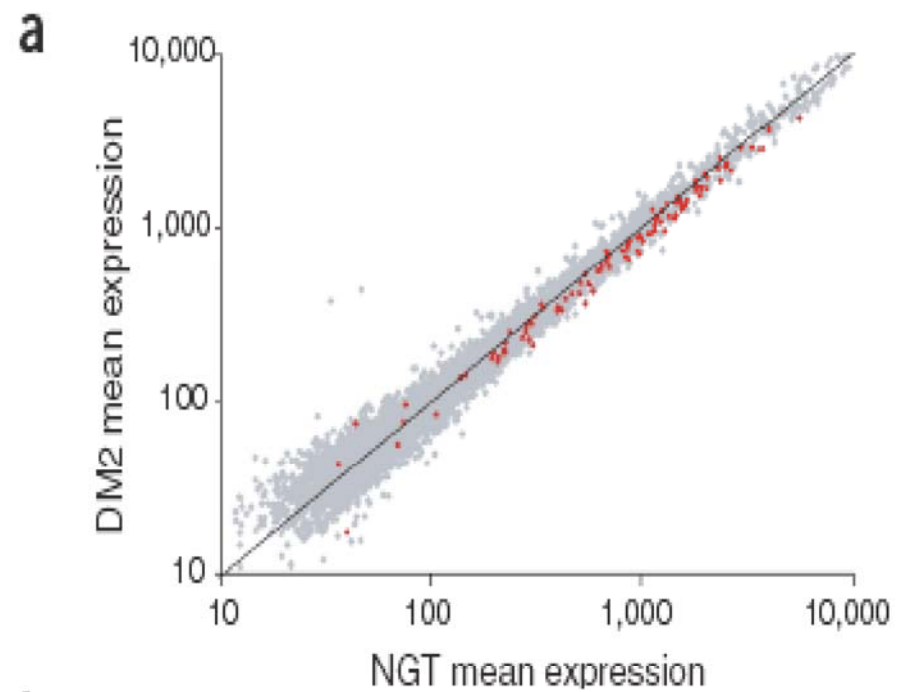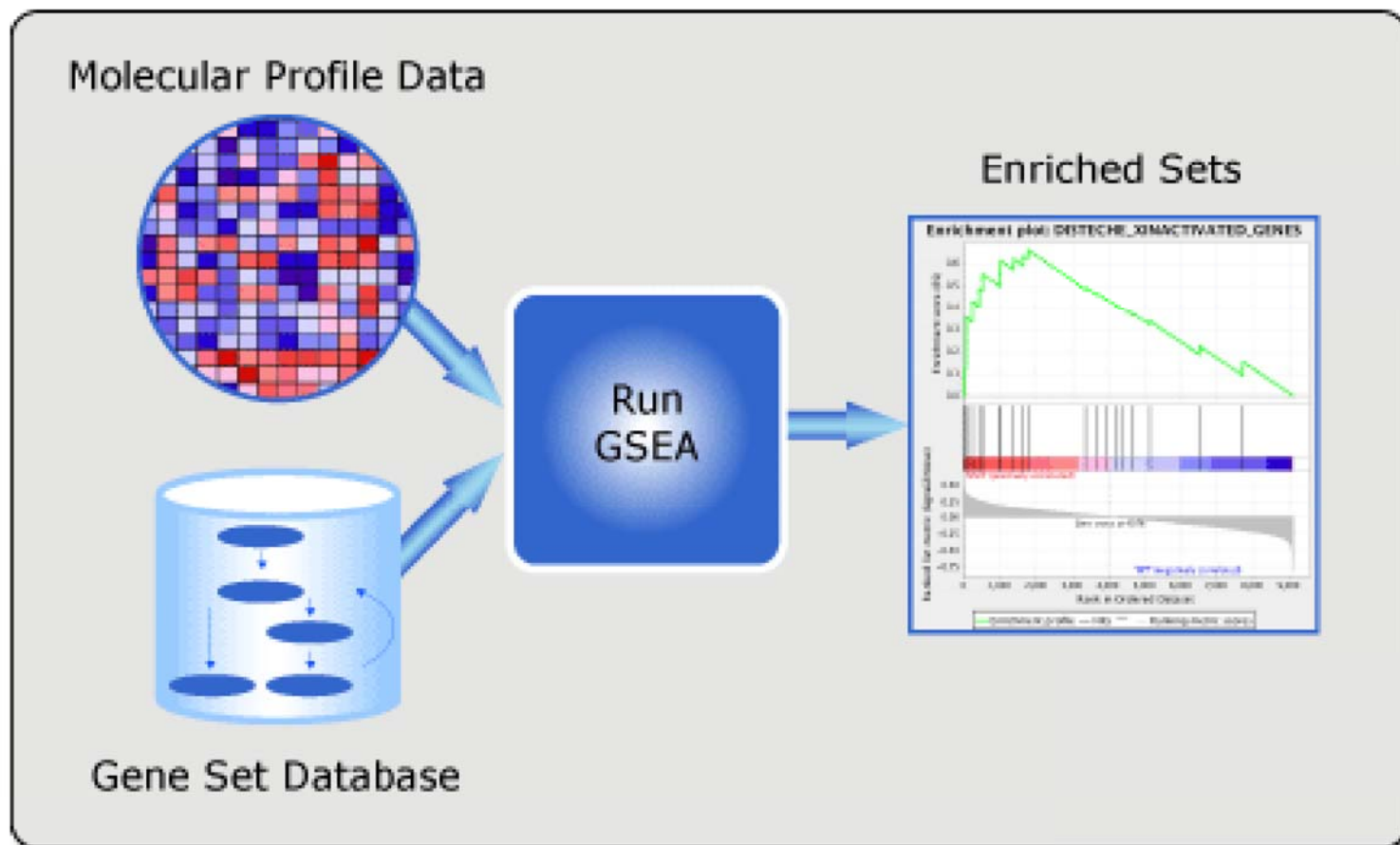- A subset of OXPHOS genes was strongly upregulated in response to PGC-1a.

1. Collect gene sets

Pathways
GO terms
Gene clusters

2. Order genes (R_i) by expression difference

NGT   DM2   Member of gene set P?

High
R_1 No
R_2 Yes
R_3 Yes
R_4 Yes
        No
        No
Low   Yes

3. Measure ES for each gene set

Running sum

ES

Genes R_1 ⟶ R_N

5. Permute class labels (1,000 times)

4. Record MES for actual data

Description	ES
WICGR OXPHOS	346
WICGR mitochondria	215
Mitochondria keyword	207
Cluster c20	181
GenMAPP OXPHOS	149
:	:
GenMAPP retinol metabolism	0

Gene set	ES
Pyruvate metabolism	138

Gene set	ES
Glycolysis	98

Gene set	ES
Cluster c29	44
Cluster c33	40
GenMAPP cysteine metabolism	22

Permutations

MES

6. Evaluate significance of actual MES against 1,000 permuted MES

9

Table 1 Clinical and biochemical characteristics of male subjects with NGT, IGT and DM2

	Class NGT	Class IGT	Class DM2	P value NGT versus IGT	P value IGT versus DM2	P value NGT versus DM2
n	17	8	18			
Age, y	66.1 (1.0)	66.4 (1.6)	65.5 (1.8)			
BMI, kg/m^2	23.6 (3.4)	27.1 (4.8)	27.3 (4.0)			5.70×10^{-3}
WHR	0.91 (0.09)	0.97 (0.04)	0.99 (0.03)	3.00×10^{-2}		3.83×10^{-3}
Triglycerides, mmol/l	1.03 (0.40)	1.83 (1.60)	2.04 (1.13)			2.63×10^{-3}
Cholesterol, mmol/l	5.39 (0.09)	4.60 (1.48)	5.77 (0.97)			
OGTT						
Glucose 0 min, mmol/l	4.67 (0.50)	5.05 (0.46)	7.83 (2.3)		9.22×10^{-5}	2.01×10^{-5}
Insulin 0 min, µU/ml	5.41 (3.3)	13.38 (8.9)	12.0 (6.0)	4.05×10^{-2}		4.10×10^{-4}
Glucose 120 min, mmol/l	6.58 (0.94)	9.15 (0.8)	14.9 (4.0)	2.51×10^{-6}	8.91×10^{-6}	4.90×10^{-8}
Insulin 120 min, µU/ml	33.5 (19.3)	125.1 (66.1)	43.5 (25.6)	5.47×10^{-3}	9.73×10^{-3}	
M value, mg/kg/min	8.74 (3.15)	6.32 (3.08)	4.22 (1.72)			2.30×10^{-5}
VO2max, ml O$_2$/kg/min	32.1 (5.46)	26.5 (4.6)	24.3 (5.6)	1.72×10^{-2}		3.09×10^{-4}
Glycogen, mmol/kg	371.1 (77.0)	326.5 (88.0)	350.6 (97.8)			
Type I fibers						
Number, %	37.2 (13.5)	33.5 (3.6)	36.4 (9.3)			
Area, %	39.1 (14.4)	32.7 (0.91)	40.1 (10.7)		2.35×10^{-2}	
Capillaries/fiber	3.91 (0.72)	4.05 (1.04)	4.14 (0.75)			
Type IIb fibers						
Number, %	73.8 (42.1)	60.2 (51.4)	72.2 (36.7)			
Area, %	31.3 (18.0)	24.7 (18.3)	36.2 (15.4)			
Capillaries/fiber	2.97 (0.71)	3.05 (0.87)	3.02 (0.65)			

Values are mean (s.d.). OGTT, oral glucose tolerance test. M value is the total body glucose uptake. VO2max is the total body aerobic capacity. Only values of $P < 0.05$ are shown for pairwise comparisons, using a two-sided t-test.

a

DM2 mean expression (y-axis): 10, 100, 1,000, 10,000

NGT mean expression (x-axis): 10, 100, 1,000, 10,000

b

Fraction of genes (y-axis): 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6

Mean expression difference (NGT − DM2) (x-axis): −400, −300, −200, −100, 0, 100, 200, 300, 400, >400

11

Molecular Profile Data

Gene Set Database

Run GSEA

Enriched Sets

Lung ADC studies

- Microarray gene expression data from tumor samples of LADC patients and their clinical outcome data, classified as "good" or "poor" outcome.

- Boston: 62 samples. Michigan: 86 samples. Stanford: 24 samples.

- BS denotes the set of top 100 genes correlated with poor outcome in Boston study. Similarly the notation MS and SS.

- The overlap is distressingly small.
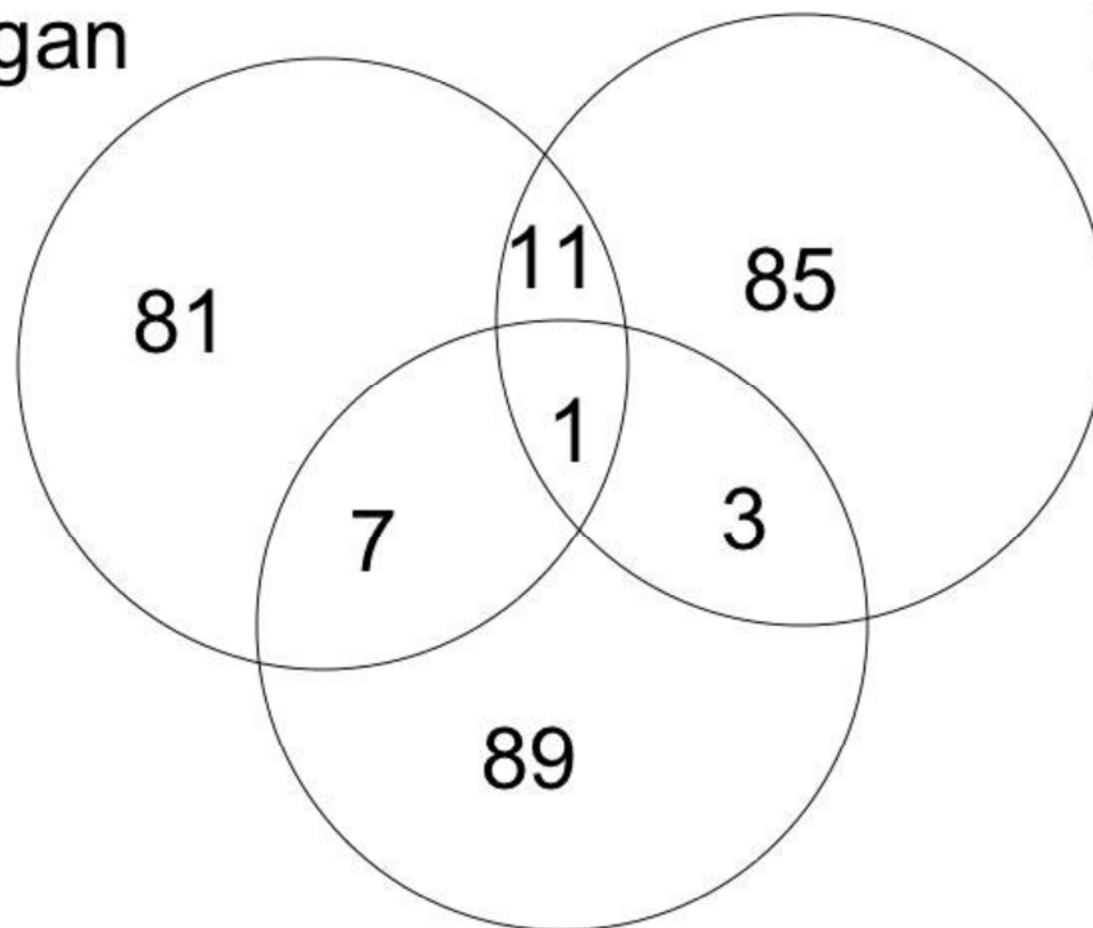
GSEA vs single-gene methods

- Looking for differentially expressed gene sets.
- Gene sets, defined by pathways or biological processes, are more reproducible and interpretable than single-gene in a large scale experiment.
- When members in a gene set exhibit strong correlation among them, GSEA is more signal-revealing.
- Leading-edge analysis helps to elucidate the results.

Enrichment Score

Given microarray data from cancer cell samples and normal cell samples. Two-sample test to look for differentially expressed genes. Let $t_i = t(g_i)$ denote the statistic for g_i. Order these genes by $t_i = t(g_i)$. Assume $L=(g_1,...,g_N)$ is ordered by t_i.

Enrichment score
(Kolmogrov-Siminov test)

Let S be a given gene set. To know if S is associated with cancer status. Define

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

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

$$ES(S) = \max\{P_{hit}(S,i) - P_{miss}(S,i) \mid i = 1, ..., N\}$$

A GSEA overview illustrating the method.
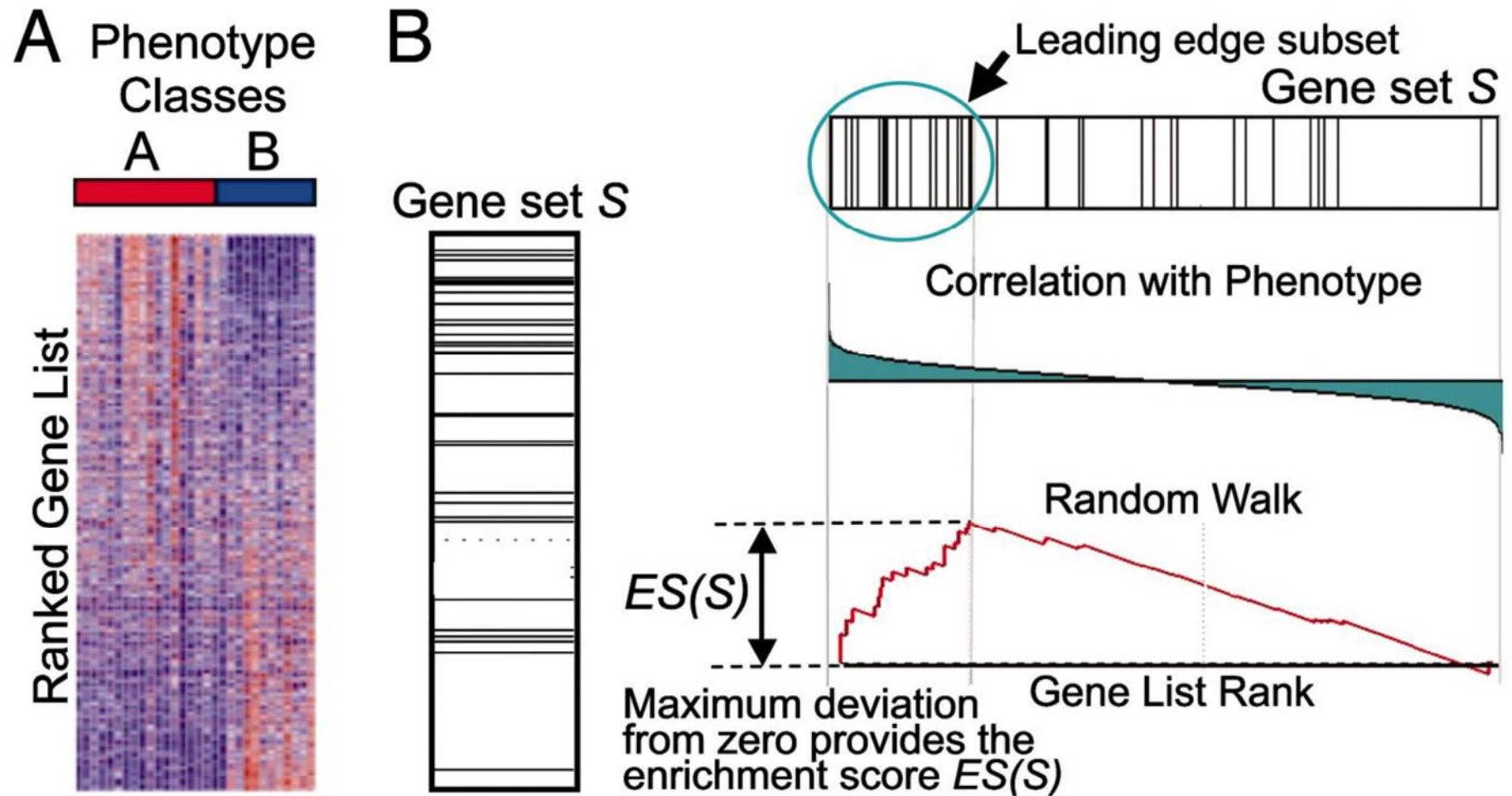
Subramanian A et al. PNAS 2005;102:15545-15550

PNAS 18

Significance of Enrichment Score

- Randomize the labels among the samples, compute the enrichment score based on this randomized sample.

- Do it 1000 times.

- Use the positive or negative portion of the null enrichment scores to get the p-value, depending on the sign of ES(S).

Multiple hypothesis testing

To adjust for the size of gene sets.

Given S and 1000 permutation, compute $E(S, \pi_k)$, for k=1,..., 1000.

If $E(S, \pi)$ is positive (negative), divide it by the mean of the positive (negative) $E(S, \pi_k)$. Denote it by $NES(S, \pi)$

Divide by the same number to get $NES(S)$.

False Discovery Rate

Given S_*, denote $NES(S_*)$ by NES^*.
Suppose it is positive.

$$\frac{\dfrac{\#\{(S,\pi)\mid NES(S,\pi) > NES^*\}}{\#\{(S,\pi)\mid NES(S,\pi) > 0\}}}{\dfrac{\#\{S\mid NES(S) > NES^*\}}{\#\{S\mid NES(S) > 0\}}}$$

Negative case treated similarly.

Lung ADC Studies

- Given the gene set BS and consider the entire ranked gene list from Michigan array data. BS has NES=1.90, p-value <0.001.

- Conversely, the gene set MS has NES =2.13, p-value<0.001.

- For GSEA using C2, at FDR 0.25, there are good overlaps.

- Greater biological insight can be obtained at lower threshold.

Boston Dataset — Gene Set: $S_{Michigan}$
Michigan Dataset — Gene Set: S_{Boston}
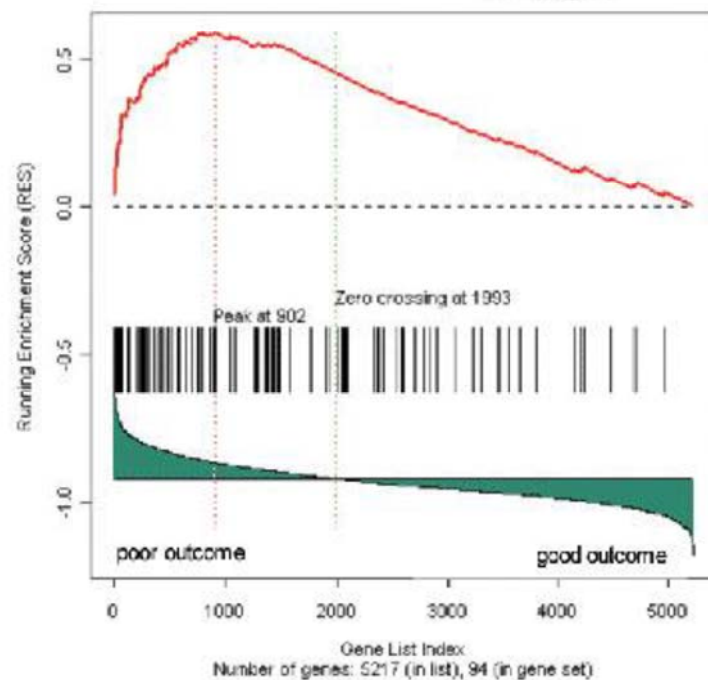
23
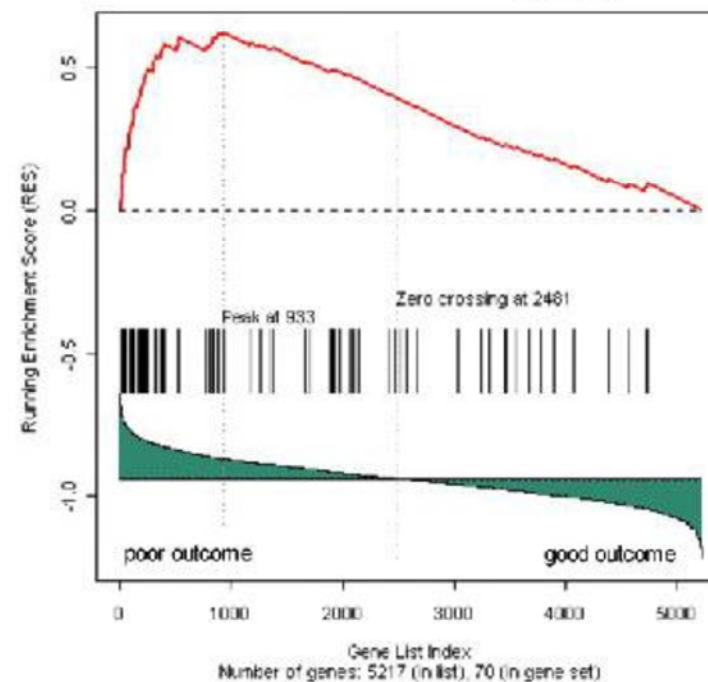
Molecular Signatures Database

- C1: Positional gene sets for each human chromosome and each cytogenetic band (326 gene sets)

- C2: Curated gene sets collected from various sources such as online pathway databases, publications in PubMed, and knowledge of domain experts. 3272 gene sets in total; CGP, 2392; CP, 880. BIOCARTA, 217; KEGG,186; REACTOME, 430 .

- C3: Motif gene sets, 836. MicroRNA targets, 221; TFT, 615.

- C4: Computational gene sets, 881.

- C5: GO gene set, 1454. Biological process, 825; Celleur components,233;MC1: Positional gene sets for each human chromosome and each cytogenetic band (326 gene sets)

Data set: Lung cancer outcome, Boston study

Enriched in poor outcome

Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216

Data set: Lung cancer outcome, Michigan study

Enriched in poor outcome

Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

- This analysis shows much greater consistency across the three lung data sets by using GSEA than by single-gene analysis. Moreover, it helps to generate compelling hypotheses for further exploration. In particular, 40 of the 60 top scoring gene sets across these three studies give a consistent picture of underlying biological processes in poor outcome cases.

- Striking evidence in all three studies of the effects of rapid cell proliferation, including sets related to Ras activation and the cell cycle as well as responses to hypoxia including angiogenesis, glycolysis, and carbohydrate metabolism. More than one-third of the gene sets (23 of 60) are related to such processes. These responses have been observed in malignant tumor microenvironments where enhanced proliferation of tumor cells leads to low oxygen and glucose levels (26). The leading-edge subsets of the associated significant gene sets include hypoxia-response genes such as HIF1A, VEGF, CRK, PXN, EIF2B1, EIF2B2, EIF2S2, FADD, NFKB1, RELA, GADD45A, and also Ras/MAPK activation genes (HRAS, RAF1, and MAP2K1).

- Strong evidence for the simultaneous presence of increased amino acid biosynthesis, *mTor* signaling, and up-regulation of a set of genes down-regulated by both amino acid deprivation and rapamycin treatment (27). Supporting this finding are 17 gene sets associated with amino acid and nucleotide metabolism, immune modulation, and *mTor* signaling. Based on these results, one might speculate that rapamycin treatment might have an effect on this specific component of the poor outcome signal. There is evidence of the efficacy of rapamycin in inhibiting growth and metastatic progression of non-small cell lung cancer in mice and human cell lines (28).

Choosing the value of p

Male vs Female Lymphoblastoid Cells

- 15 males and 17 females to identify gene sets correlated with the distnction "M>F" and "F>M".

Data set: Lymphoblast cell lines	
Enriched in males	
chrY	<0.001
chrYp11	<0.001
chrYq11	<0.001
Testis expressed genes	0.012
Enriched in females	
X inactivation genes	<0.001
Female reproductive tissue expressed genes	0.045

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

Original (4) enrichment score behavior.
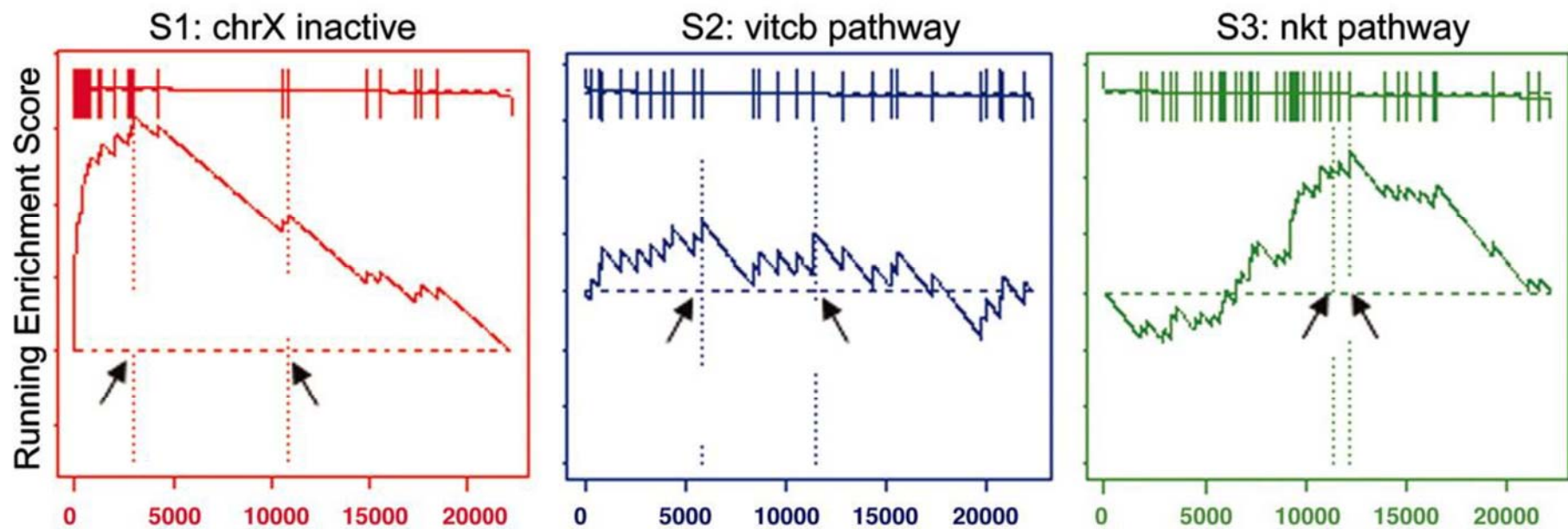
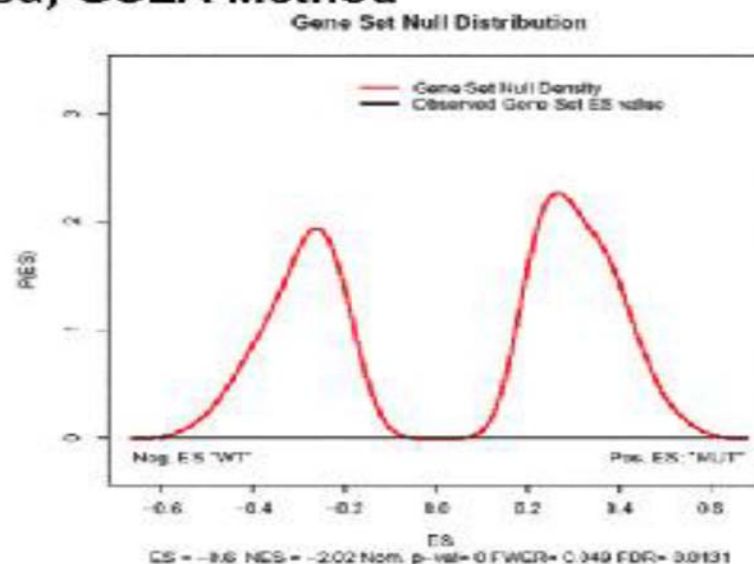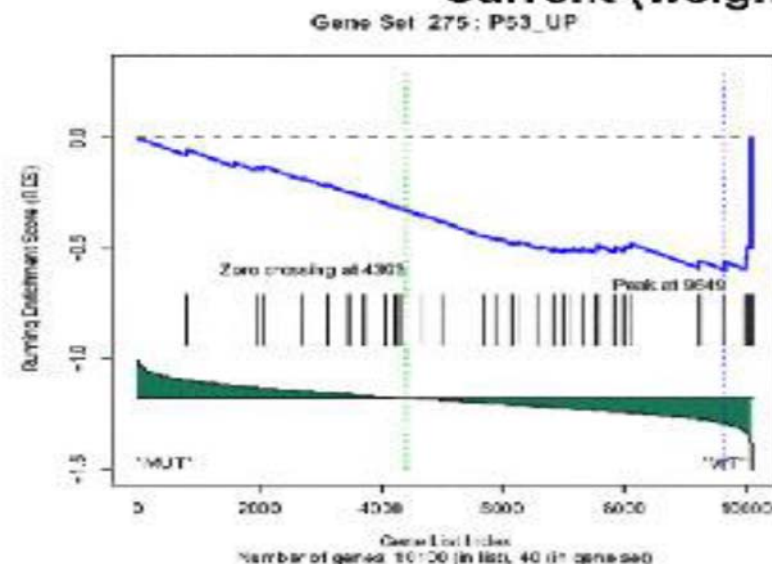# p53 status in cancer cell lines

- 17 are normal and 33 carry mutations in p53, among NCI-60 cell lines.

p53 Upregulated Genes in p53 Wild Type Phenotype

Original (unweighted) GSEA Method
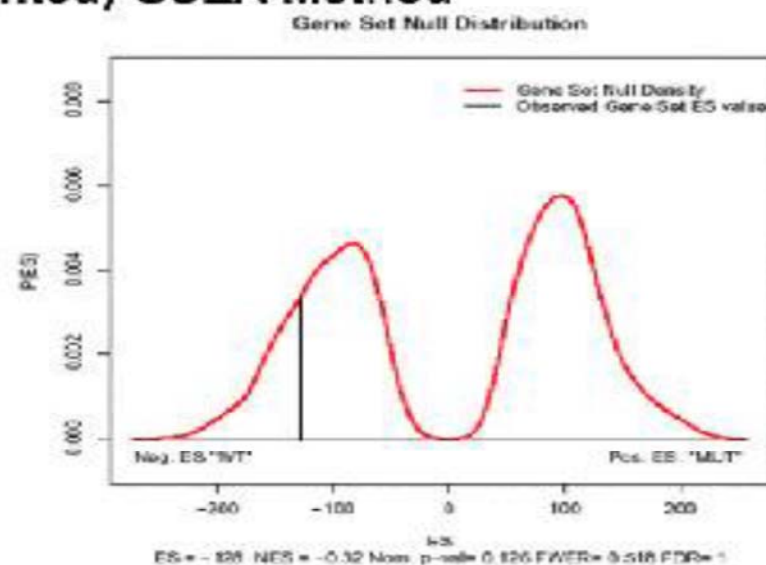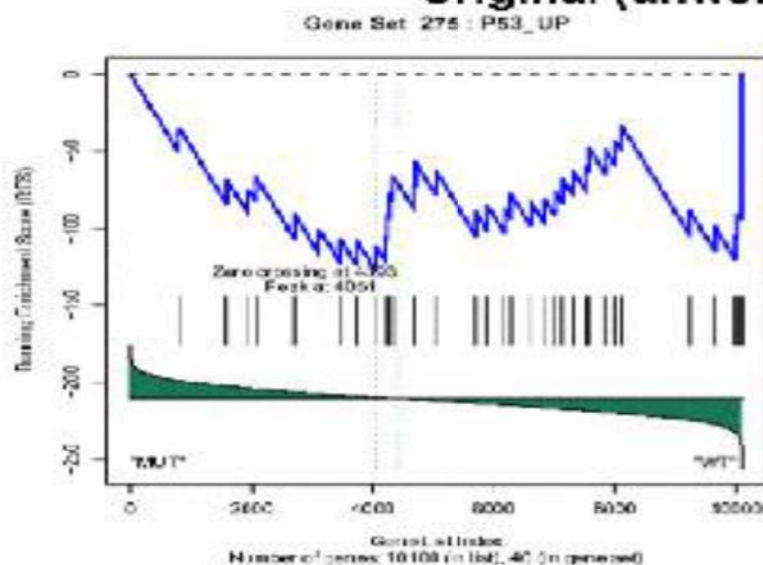

Current (weighted) GSEA Method

36

Data set: p53 status in NCl-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

Enriched in P-53 mutant

- Core genes.

- Only Ras signaling pathway is significant at FDR .25.

- The next significant ones are Ngf and Igf1 signaling pathways.

- They share core genes, component of MAPK.

Leading edge overlap for p53 study.
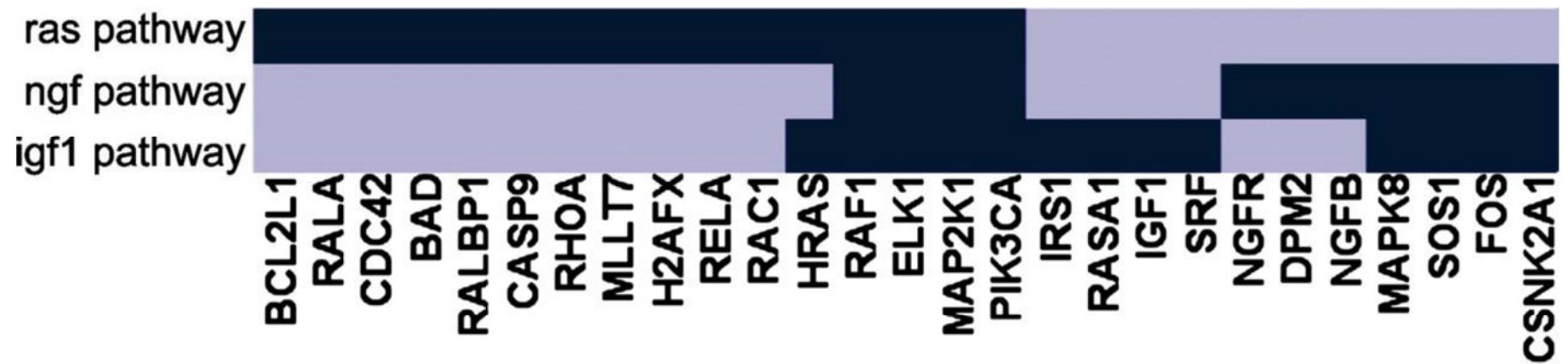
Many pathways:
Leading edge analysis

Remarks

- Kolmogorov-Smirnov type statistic. The value of p. p=1 is default. Larger p penalize genes having small correlation with the phenotype.

- Many pathways, leading edge analysis.

- The permutation method based on label shuffling is computational demanding but more biologically relevant and may produce larger FDR than other simple method. Tamayo et al. (2012)

- Over-representative methods (IPA). Aggregation methods. Third generation methods. IPA.

More systematic approaches

- Genomic data from experiments and exogenous functional information from projects like GO and KEGG.

- Measurements are at the gene level; inference is at the functional category level.

- Leading analysis implies that these pathways are related.

- Quantitative responses.

- Both expression data and SNP data.

- Efron and Tibshirani (2007) ON TESTING THE SIGNIFICANCE OF SETS OF GENES, Annals of Applied Statistics

- Newton, He, Kendziorski (2012) A model-based analysis to infer the functional content of a gene list. Stat Appl Genet Mol Biol.

Thank you