Discovering relevant pathways in microarray experiments – "honest" testing methods

Fred A. Wright and Zhen Li

Department of Biostatistics and Carolina Environmental Bioinformatics Center

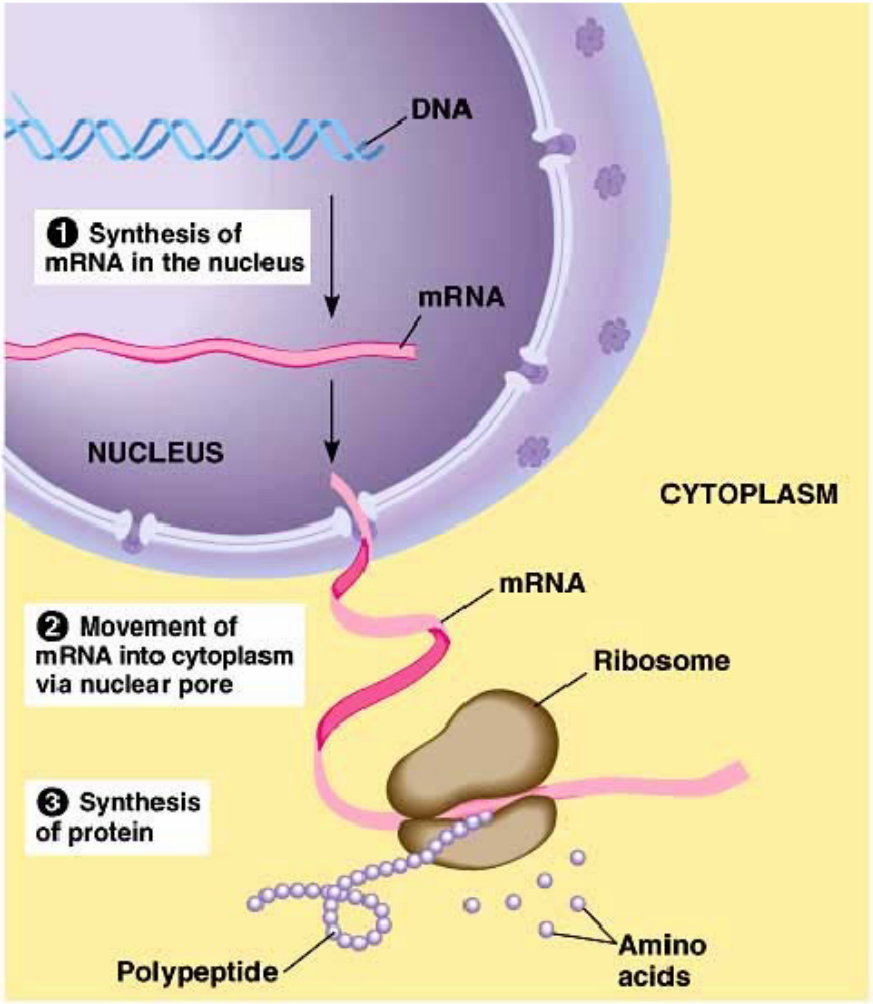
University of North Carolina at Chapel Hill

UNC SCHOOL OF PUBLIC HEALTH

Gene category testing methods

- Test for *enrichment* of a known pathway, gene ontology or other functional keyword among list of “significant” genes.
- These techniques have been used in other talks in the Info on Informatics series
- The underlying principle is simple, and similar techniques will become even more common
- We will refer generically to the keyword as a *category* to which a gene may belong

The central dogma of molecular biology

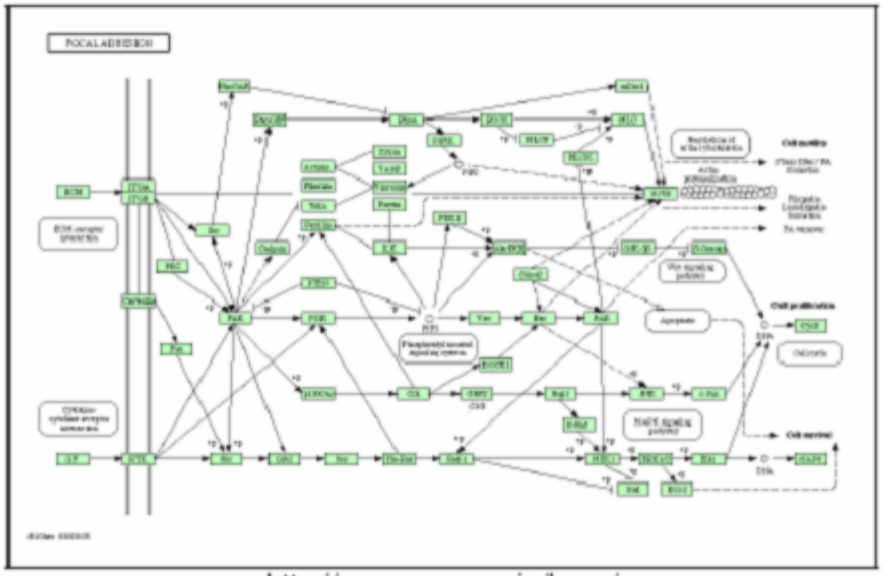


http://www.khugene.com/

1. DNA
   - SNP arrays: genotype studies
   - array CGH: DNA copy number
   - ChIP-chip: TF binding sites
2. mRNA
   - DNA microarrays: cDNA, Affymetrix, Agilent
   - SAGE
3. Proteins
   - 2-D electrophoresis
   - Maldi-Tof mass spec

- GOStats, GOMiner, GOSurfer, GO Tree Machine are examples oriented to Gene Ontology (see reference list), and several other software packages perform similar analysis (SAFE, EASE, GSEA).
- Many commerical expression analysis packages do some form of category testing, and some packages are heavily based on it (PathArt, Ingenuity).

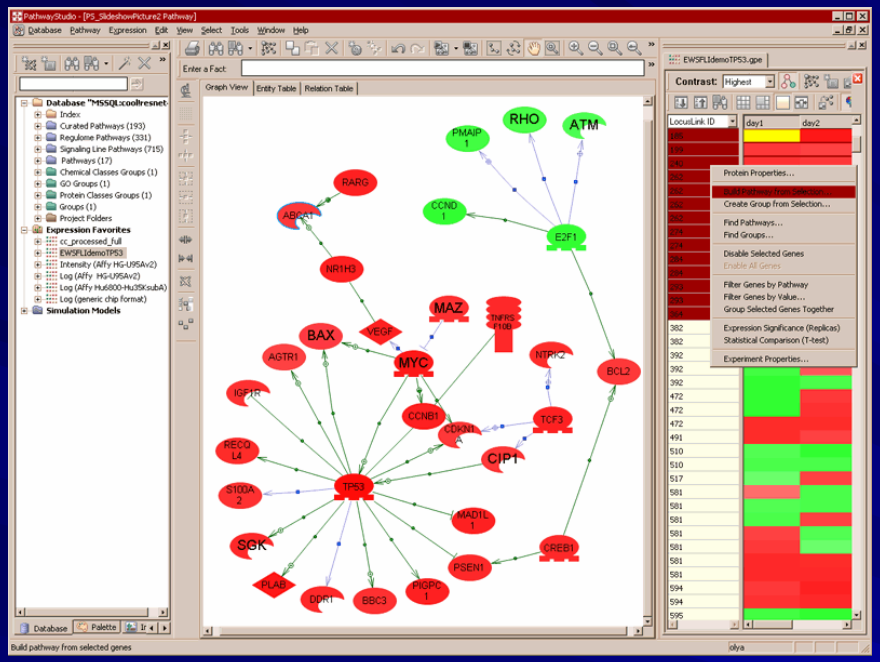
- Ready availability of comprehensive annotations of known genes, and the probe(set)s of different microarray platforms:
  e.g., Gene Ontology (GO), KEGG, TRANSFAC



http://www.genome.jp/kegg/

- We look at aggregate behavior within a category
- Hope to identify patterns, possibly moderate but consistent gene-specific effects
- Potentially reduce the number of hypothesis tests (10X or more)

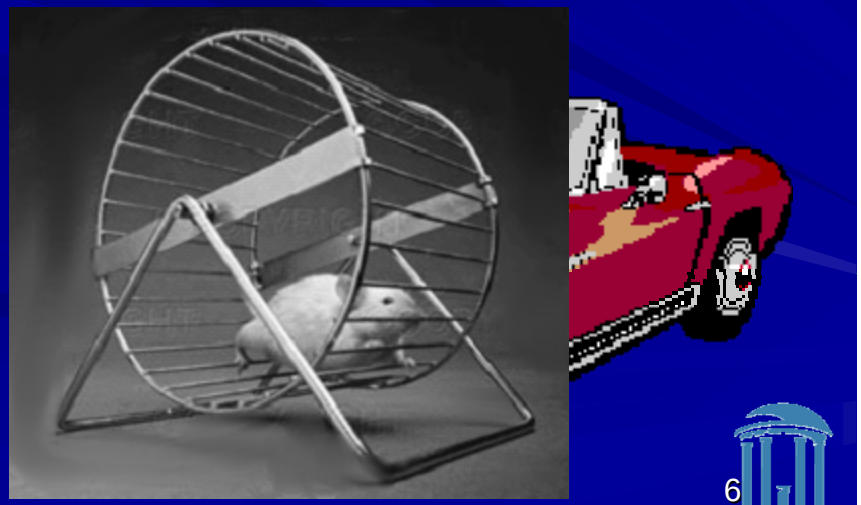
Pretty Pictures….



http://www.ariadnegenomics.com

Make us feel like we understand….

But what's under the hood?

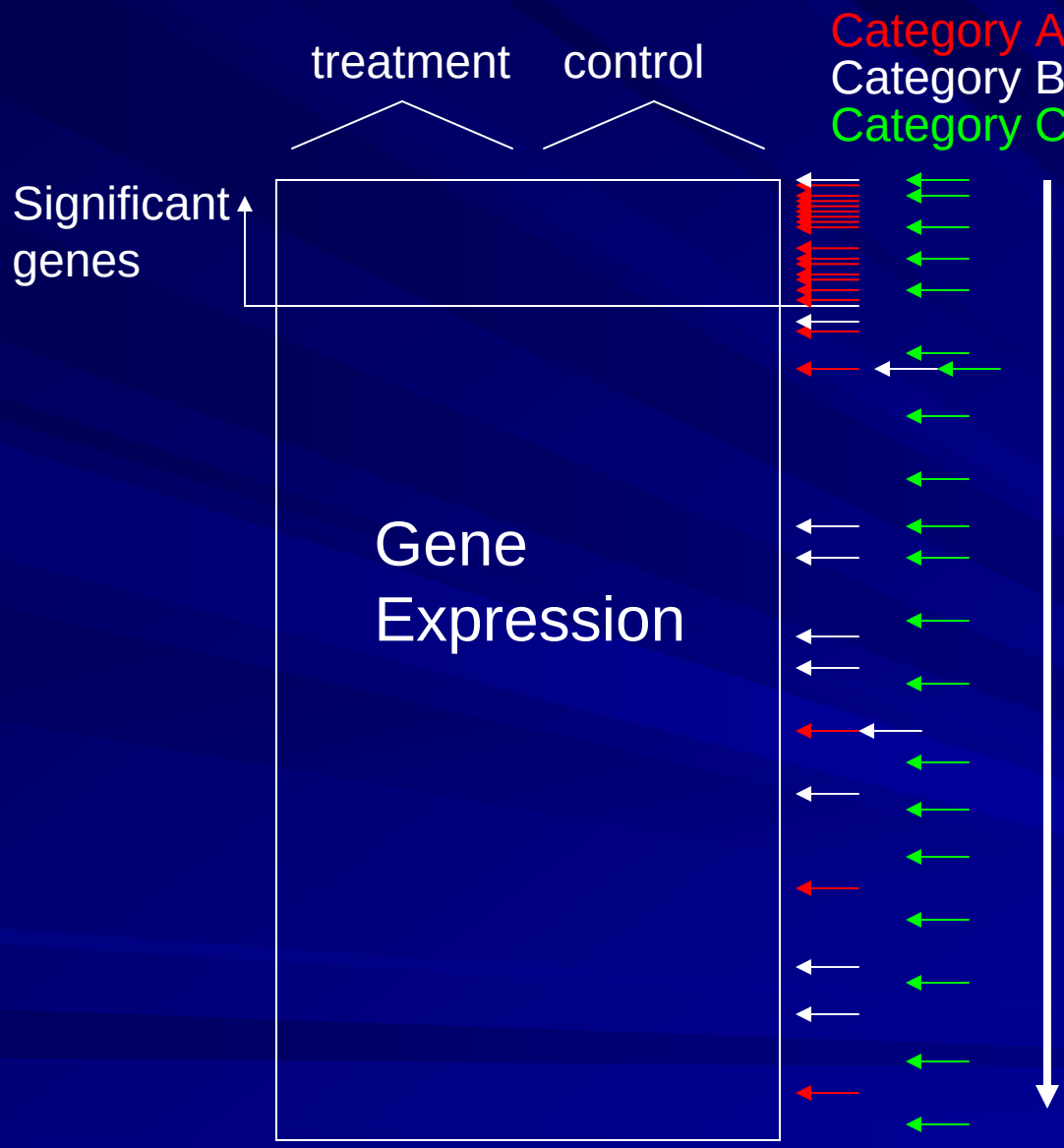


“Gene-list” methods

	Not in category	In category	
Not significant	9485	15	9500
significant	495	5	500
	9980	20	

Fisher’s exact test for enrichment (one sided) p=0.00254

Binomial approximation test for enrichment (one sided) p=0.00257



Category A (signif?)
Category B (not signif)
Category C

Genes ranked by differential expression

(most significant to least significant)

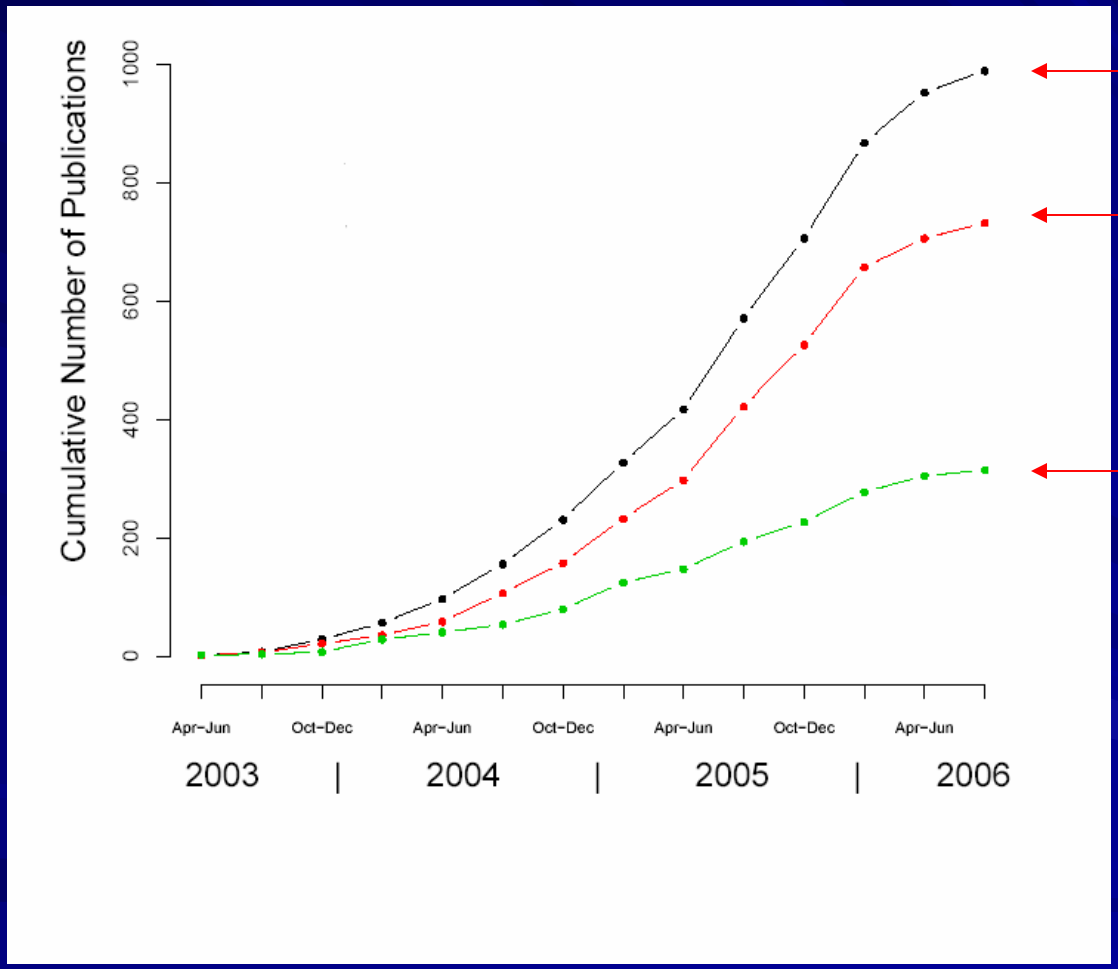
“Gene-list” methods

- Each category/pathway/keyword has an associated enrichment p-value
- There are as many such p-values as there are categories
- The set of category p-values can be subjected to standard conservative methods for controlling error rates, provided they are true p-values.

Why do we need new methods?

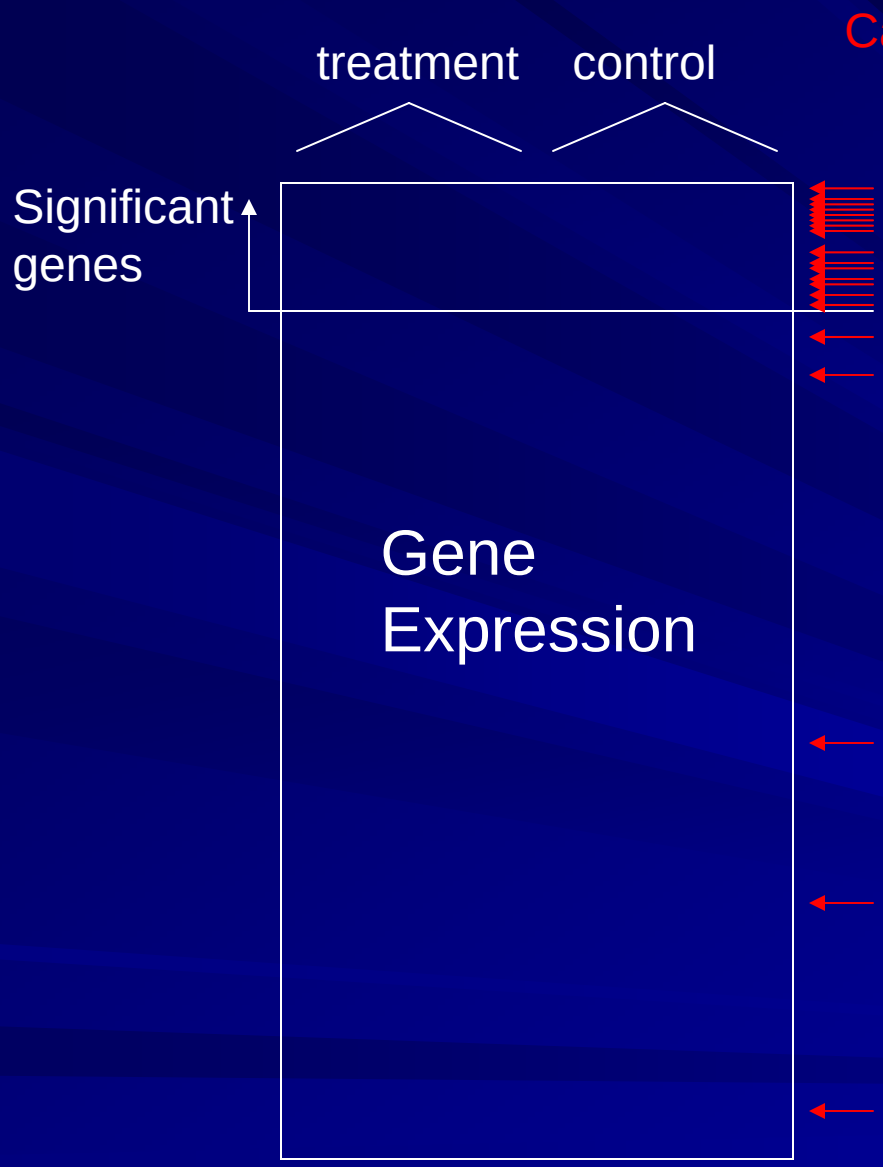
- Problem: we should use the positive correlation of categories to our advantage, reducing effective number of tests performed
- Answer: if we use permutation of arrays/samples and save our results, we can appropriately account for the correlation

Number of publications using category gene enrichment methods



Why do we need new methods?

- Problem: the gene list size is arbitrary, and doesn't rank genes within the gene list.
- Answer: we can use an overall statistic for the category that considers gene-specific p-values in a graded, continuous manner. One approach is based on the ranks of the p-values for the genes within the category vs. the remaining genes (Wilcoxon statistic).



Category A (signif)

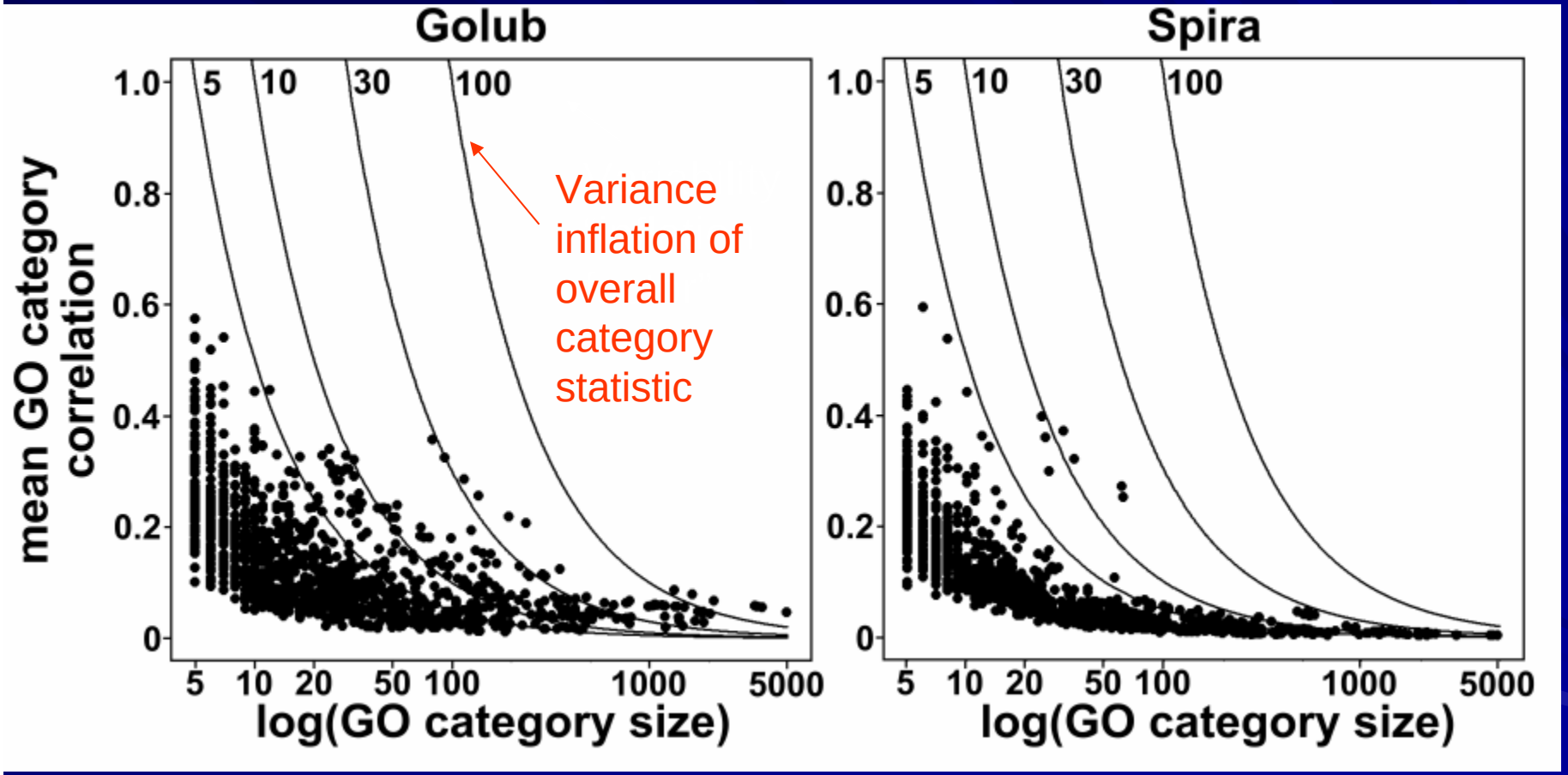
We could count the number of genes within the category that appear within the significant gene list (Fisher's test, Binomial proportions)

Or we could look at the positions of all the red arrows, look for tendency of genes to rank highly (Wilcoxon test)

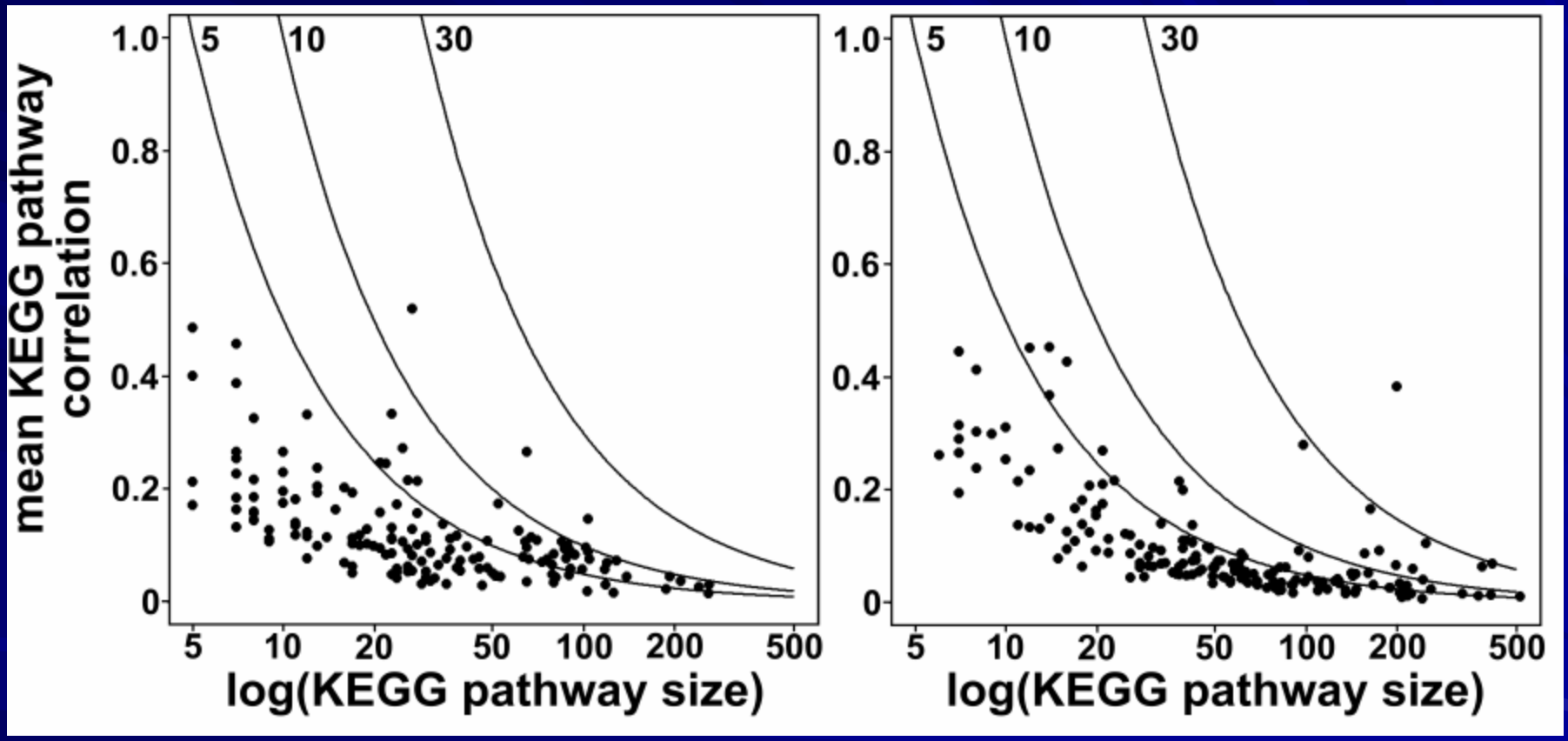
Why do we need new methods?

- Problem: standard “p-values” for keyword/category enrichment can greatly inflate false positives (Type I error)!
- First we need to understand why. Genes that are positively correlated within a category tend to be either significant or not significant *together*. This adds extra variability to the enrichment table, even under the null hypothesis.

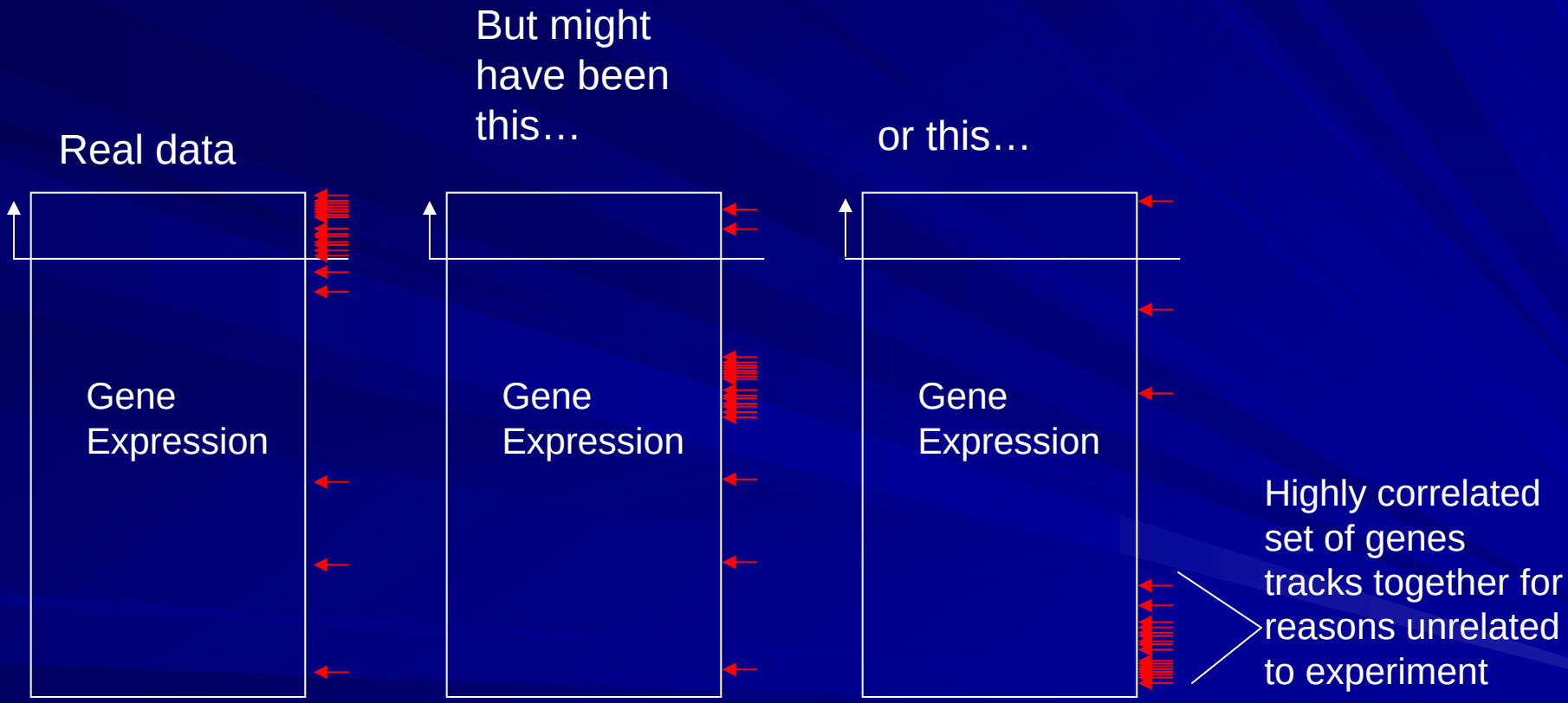
Two example datasets – internal gene correlation greatly inflates Type I error



Two example datasets – internal gene correlation greatly inflates Type I error



Why is correlation a problem?



•We have documented that simple gene list methods (assuming independence of genes) can have overall false positive rates of *75% or more*, even when applying Bonferroni correction to all categories

•Array permutation has no such inflation, and likely produces higher-quality category list. Early examples:

Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics

SAFE

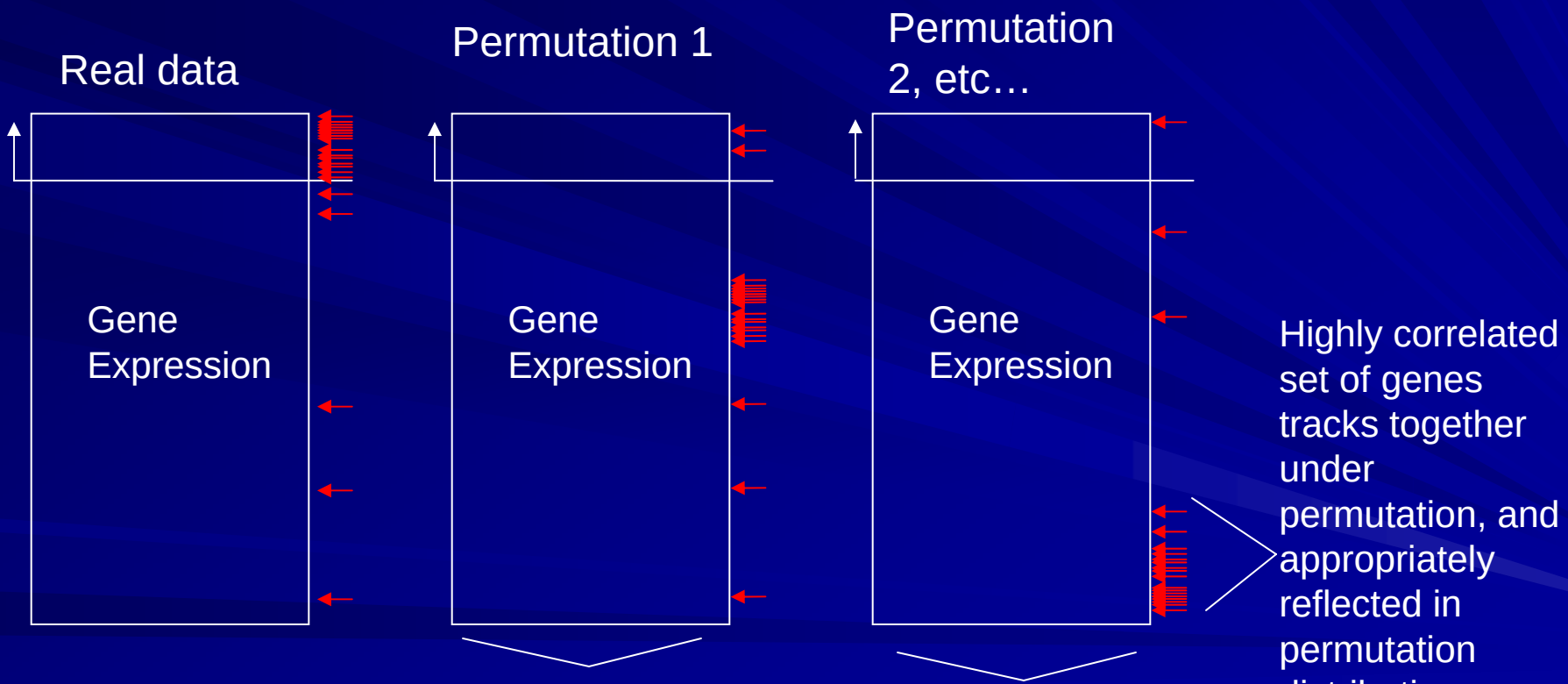
Proc. Natl Acad. Sci. USA, **98**, 1124–1129.

PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes

GSEA

Nat. Genet., **34**, 267–273.

How does array permutation address this?



The SAFE procedure

Significance Analysis of Function and Expression

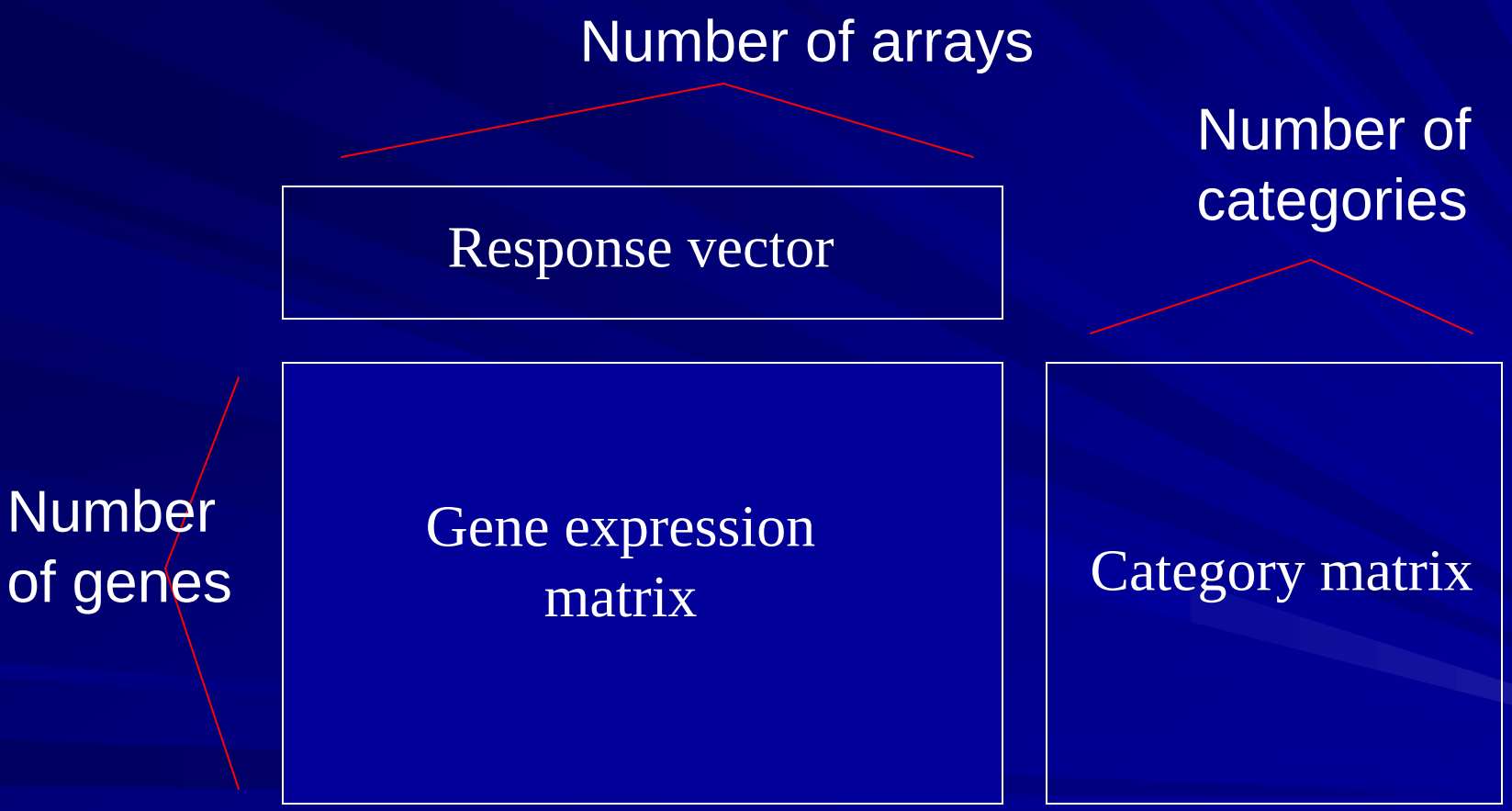
Barry, Nobel and Wright (2005) *Bioinformatics*, **21**:1943-1949.

- Extends the work from Virtaneva et al. (2001), which is essentially the same as GSEA from Mootha et al. (2003).
- Define a *response* vector of values associated with the arrays, such as disease status, survival, etc.
- Define a gene-specific *local* statistic that compares gene expression to the response (e.g., *t*-statistic)
- Define a *global* statistic that is sensitive to a category being generally more significant than other categories (e.g., the Wilcoxon rank-sum statistic for the ranks of local statistics)

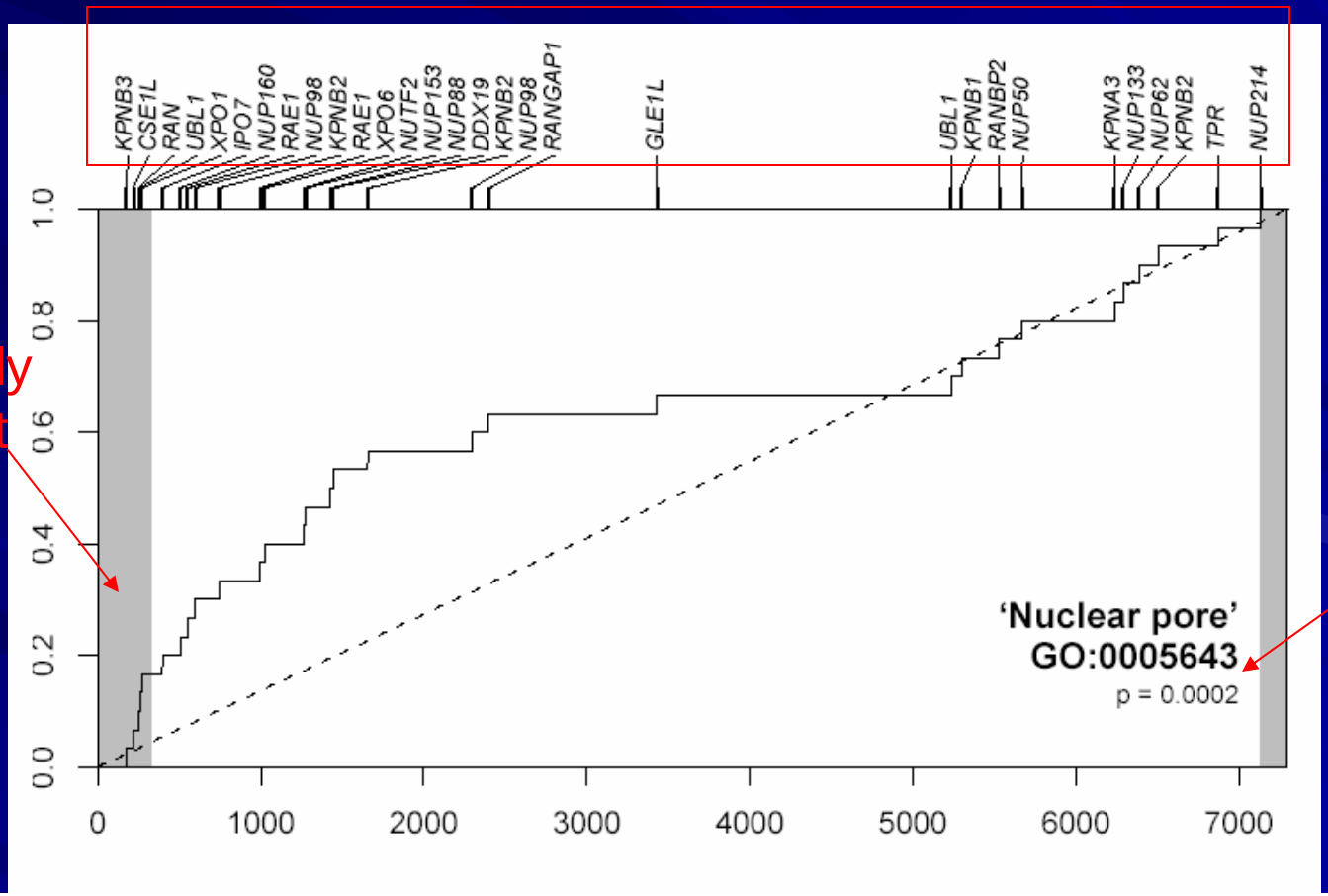
The SAFE procedure, cont.

- Define a *category* matrix that indicates for each gene *i* and each category *j* whether or not the gene belongs to the category (1=yes, 0=no).
- The category matrix can consist of any attributes. Choices may include KEGG pathways, GO, Pfam motifs, etc.
- Permute the response vector many times in order to obtain permutation-based *p*-values for each category individually, and to estimate the error rates associated with multiple category tests.

SAFE Input


Number of arrays
Number of categories
Response vector
Number of genes
Gene expression matrix
Category matrix

SAFE plots show the empirical cumulative distribution function of the ranks of the local statistics within the category. Departures from the unit line are of interest. A survival analysis example is given below for *n*=125 adenocarcinomas, with the scaled Cox regression coefficient as the local statistic.



Genes in category

Shading indicates individually significant genes

Category p-value

SAFE output of significant categories, Bhattacharjee data.

Category ID and name	Category size	*p*-value	$\widehat{\text{FDR}}$
Normal versus cancer			
GO:0016460, 'Myosin II'	10	0.0004	0.066
GO:0000786, 'Nucleosome'	19	0.0004	0.066
Pfam:PMP22_Claudin	11	0.0005	0.066
ANOVA among subtypes			
GO:0007010, 'Cytoskeleton org. and biogen.'	128	0.0003	0.064
GO:0007017, 'Microtubule-based process'	67	0.0005	0.064
GO:0006996, 'Organelle org. and biogen.'	153	0.0005	0.064
GO:0016043, 'Cell org. and biogenesis'	283	0.0007	0.064
GO:0009117, 'Nucleotide metabolism'	82	0.0008	0.064
GO:0007028, 'Cytoplasm org. and biogen.'	175	0.0011	0.087
GO:0006164, 'Purine nucleotide biosynth.'	45	0.0016	0.099
Survival of adenocarcinomas			
GO:0005643, 'Nuclear pore'	30	0.0002	0.034
GO:0046930, 'Pore complex'	30	0.0002	0.034

False Discovery Rate – accounts for number of categories tested

Current work and future directions for array permutation procedures

- Bootstrapping (different from permutation) turns out to be more powerful when many genes are differentially expressed by treatment condition
- Working on more user-friendly graphics tools

- Extensions to identifying known transcription factor motifs associated with response. Here the “category” is a probabilistic score from 0 to 1 for all genes, representing likelihood of containing the motif.

Representation of binding-site motifs

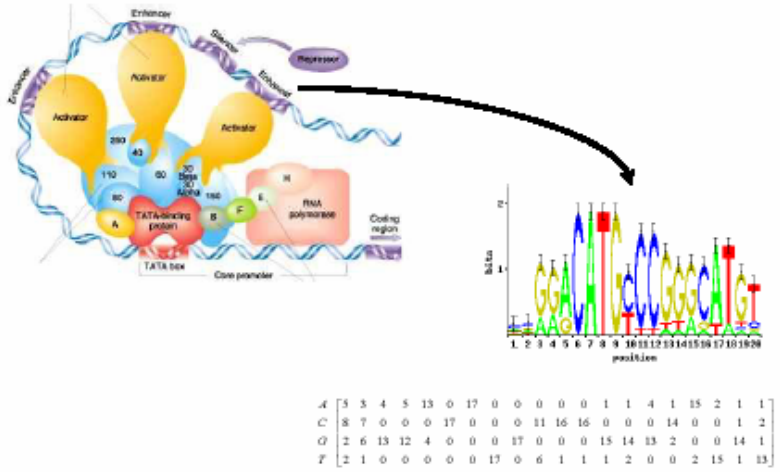


Figure: Schematic of a TF complex, the PSWM for p53 from 17 identified binding sites, and the corresponding sequence logo.

TFs are typically parametrized in a position-specific manner using a Product-multinomial: $\Theta = (\theta_1, \ldots, \theta_W)$

and a multinomial or Markov chain model for background, θ_0.

Toxicity case studies of SAFE category analysis using the SAFE package in R

- We mainly apply GO, KEGG and Pfam annotation, because these are available in R for major array platforms
- Results are still being interpreted, biology always takes some thought
- But – with lower error rates we hopefully have saved some postdocs from unnecessary followup of false leads!

SAFE Data Analysis

Purpose of the Analysis:

To find significant pathway categories in a dose-response study

Data Source: Kevin Crofton and Josh Harrill

Microarray: Affymetrix Rat Genome 230 2.0 Array

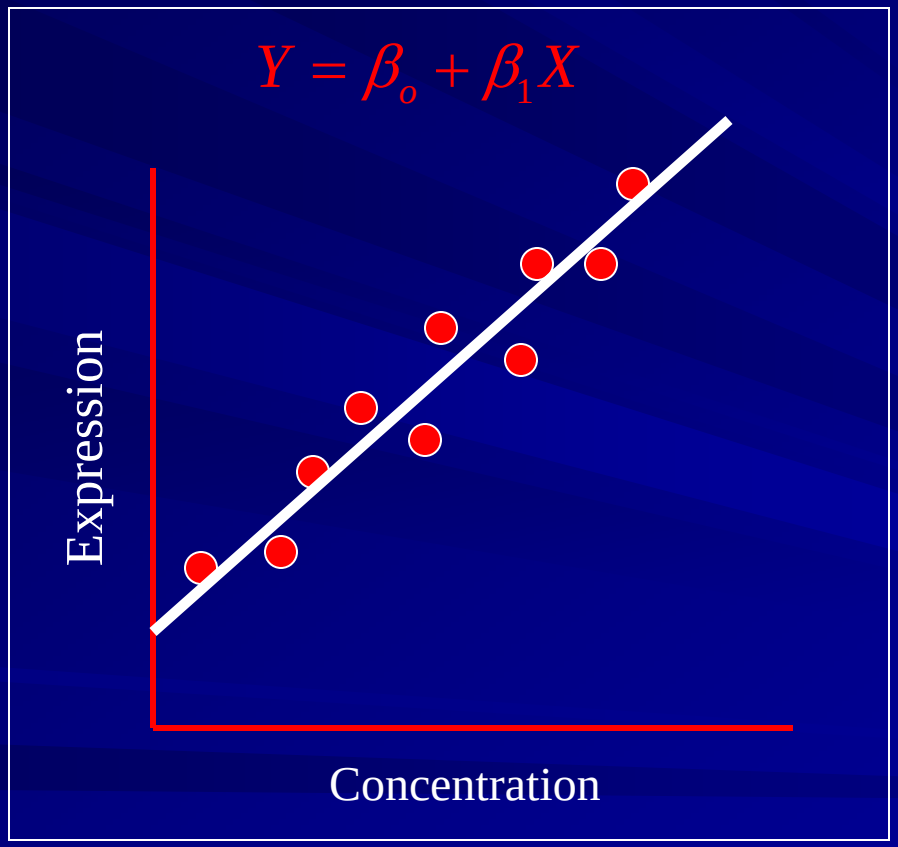
Chemicals: Deltamethrin and Permethrin

Permethrin: Vehicle Controls, 1.0 mg/kg, 10.0 mg/kg and 100.0mg/kg

Deltamethrin: Vehicle Controls, 0.3 mg/kg, 1.0 mg/kg and 3.0 mg/kg

Local Statistics: t Statistics

Simple linear regression for dose response (computationally efficient)



To test whether the slope is significantly different from 0

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

t statistics:

$$t_{(n-2)} = \frac{\beta_1}{S_{\beta_1}}$$

Global Statistics: Two Sample Binomial Proportion Test

Aim: to measure the difference between the local statistics in a category and the local statistics in the complement of that category

$$z = (\hat{p}_1 - \hat{p}_2) / \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$\hat{p}_1$: the proportion of significantly expressed genes in a category

$\hat{p}_2$ the proportion of significantly expressed genes in the complement of that category

Significant Categories Identified by SAFE

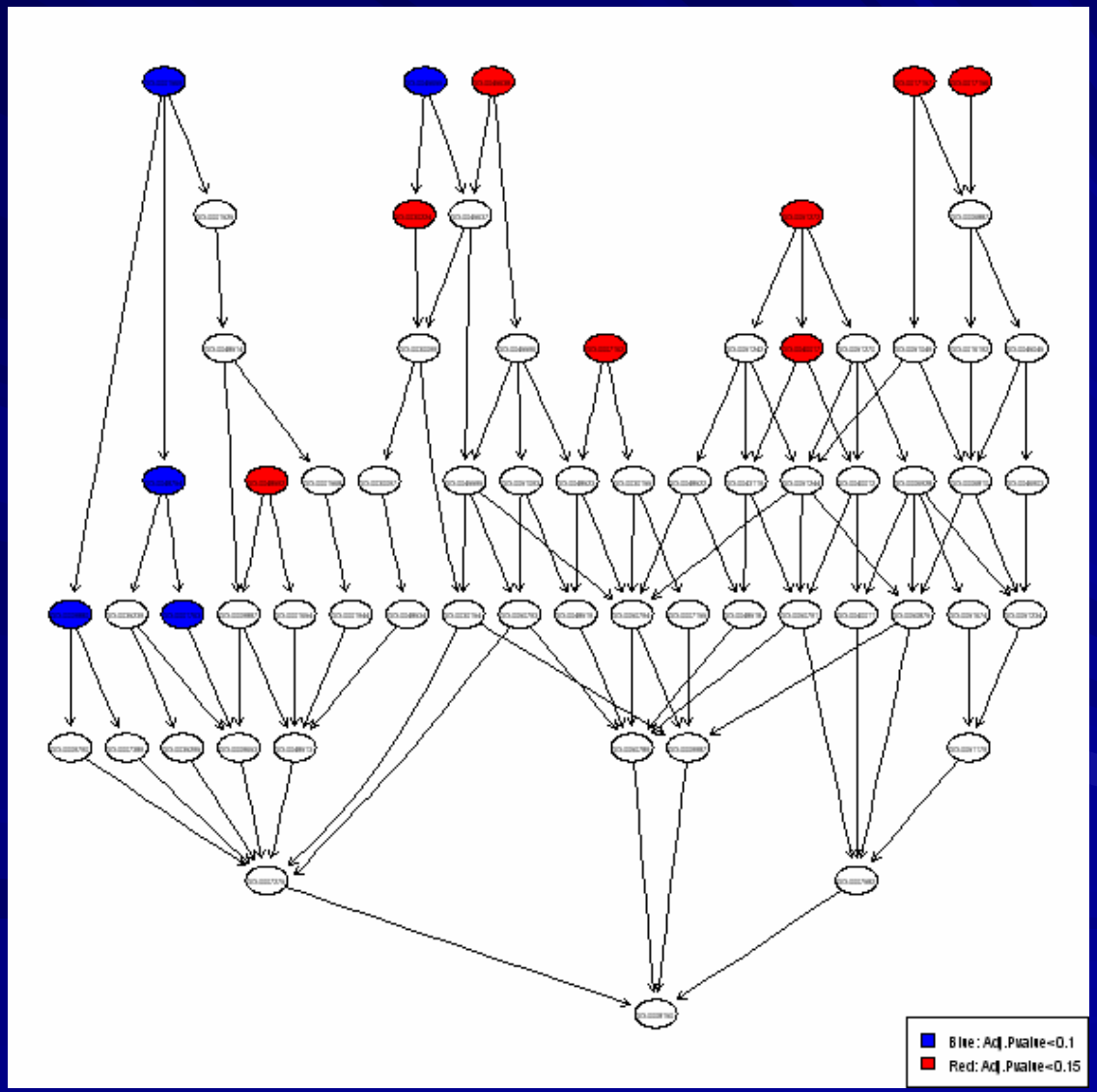
PERMETHRIN:

Categories	Term	Size	Emp.pvalue	Adj.pvalue
GO:0048754	branching morphogenesis of a tube	66	2.00E-04	0.03493
GO:0001763	morphogenesis of a branching structure	67	2.00E-04	0.03493
GO:0001569	patterning of blood vessels	31	3.00E-04	0.040601
GO:0009880	embryonic pattern specification	49	5.00E-04	0.055397
GO:0045655	regulation of monocyte differentiation	32	0.001	0.093149
PFAM:05210	Sprouty protein (Spry)	9	1.00E-04	0.050024

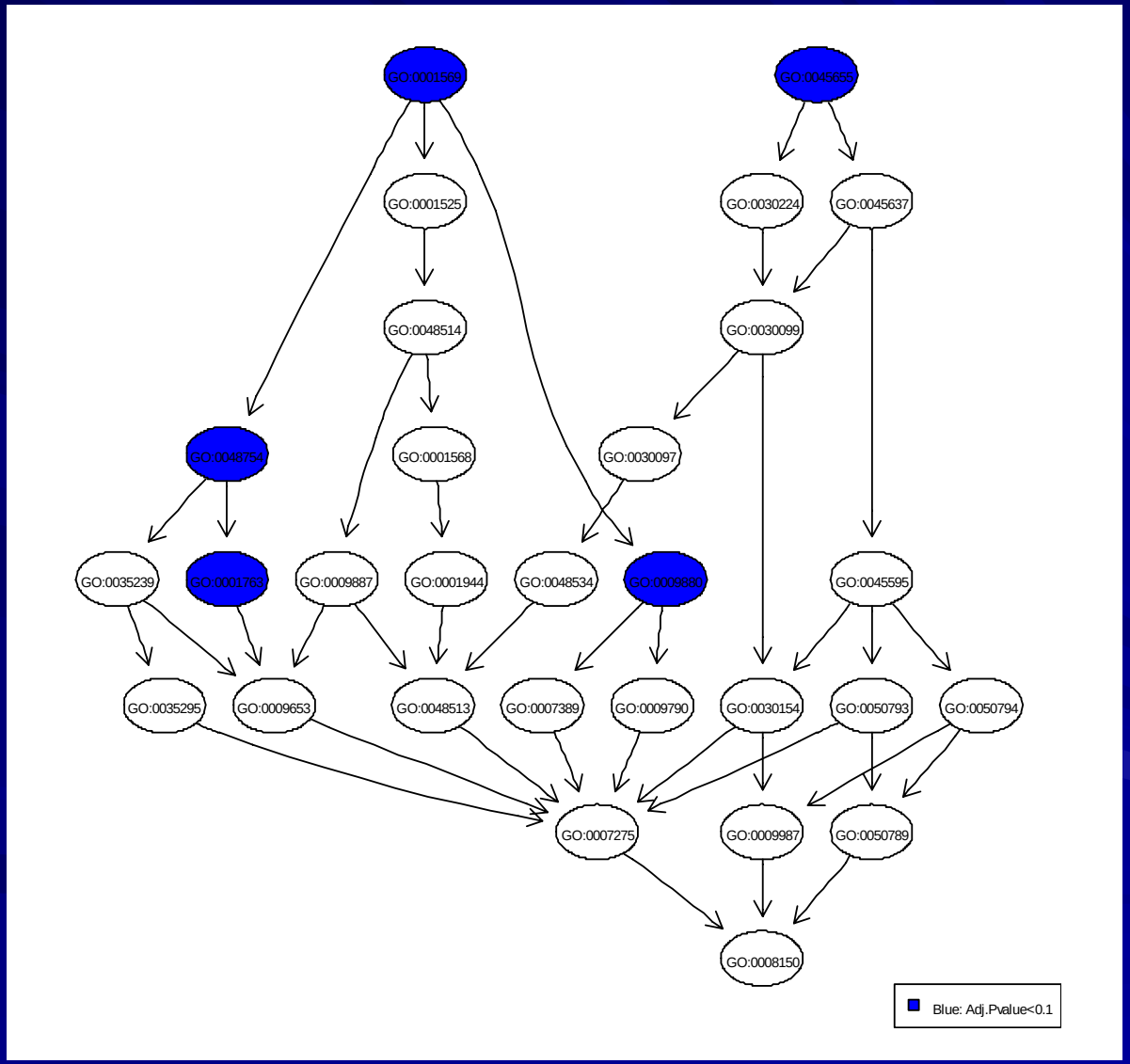
DELTAMETHRIN

Categories	Term	Size	Emp.pvalue	Adj.pvalue
KEGG:00564	Glycerophospholipid metabolism	73	7.00E-04	0.040453
KEGG:00400	Phenylalanine, tyrosine and tryptophan biosynthesis	12	0.0024	0.092856

GO BP (Biological Process) “Interesting” Categories



GO BP (Biological Process) “Interesting” Categories



SAFE Report: Detailed Pathway Information for Significant Category

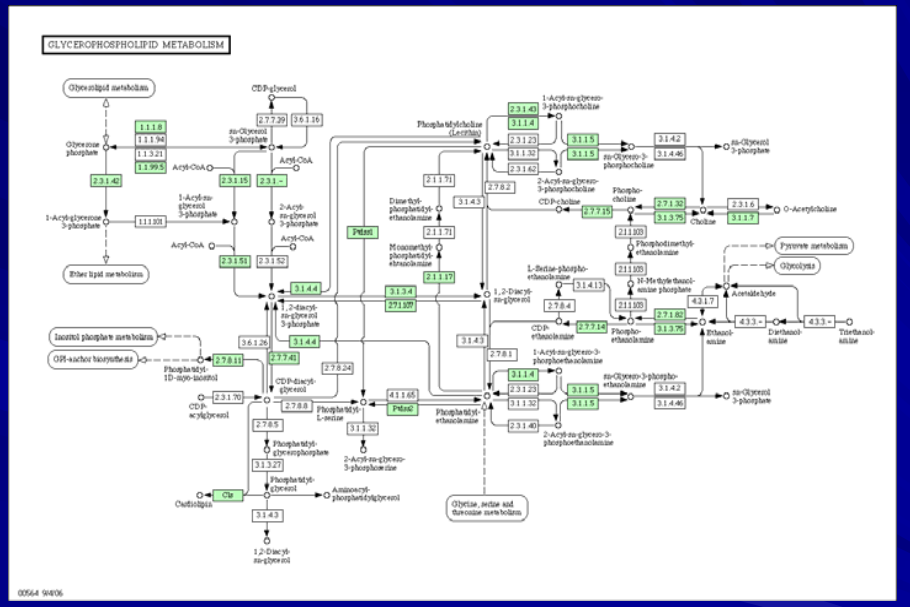
KEGG:00564 consists of 73 genes

Upregulated Genes

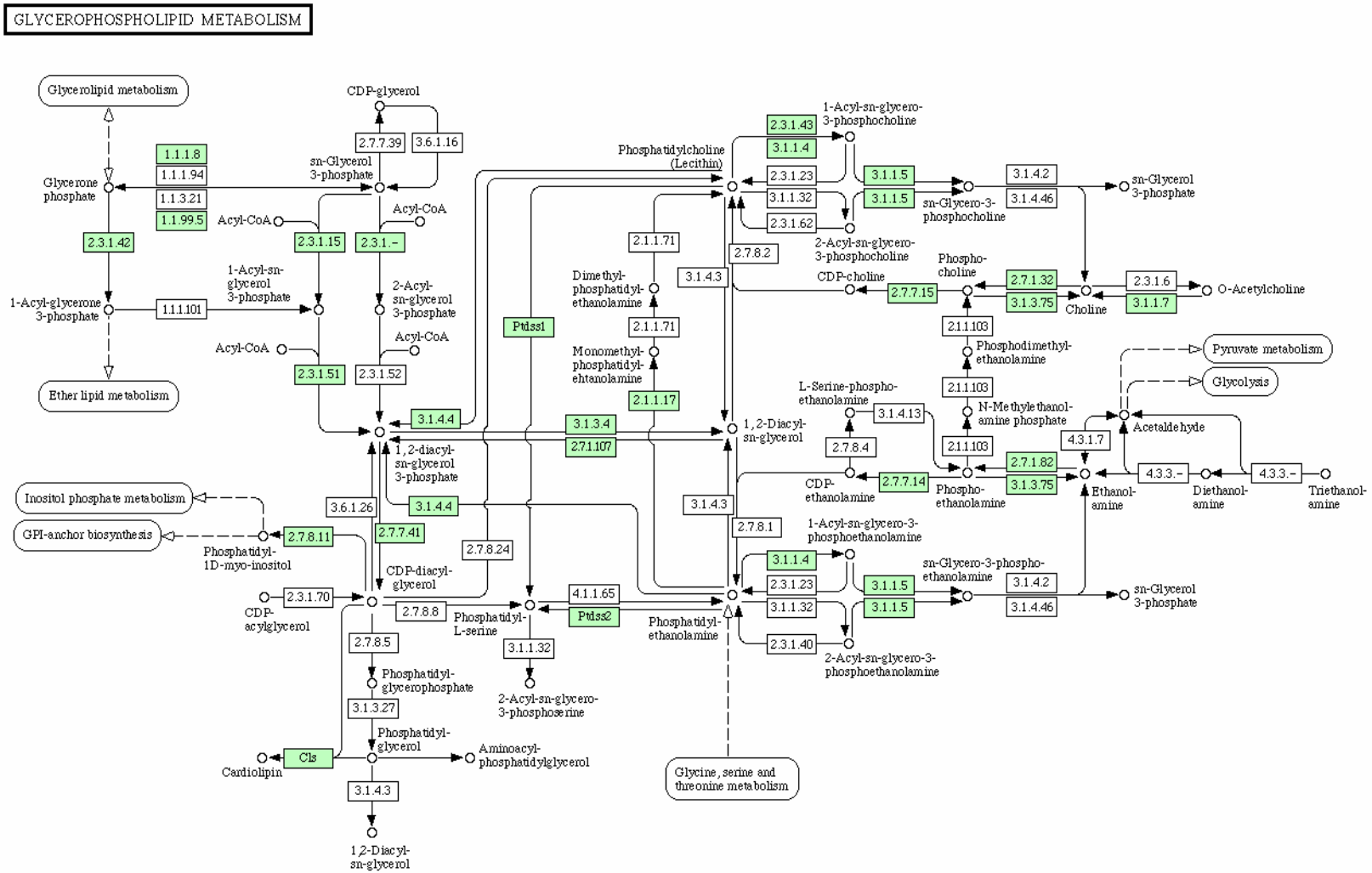
	Local.Stat	Emp.pvalue
1371363_at	5.640	2e-04
1369560_at	4.550	2e-04
1368891_at	3.113	0.0039
1387265_at	2.295	0.0269
1382772_at	1.962	0.0588
1370385_at	1.614	0.1133
1374109_at	1.347	0.1809
.		
.		

Downregulated Genes

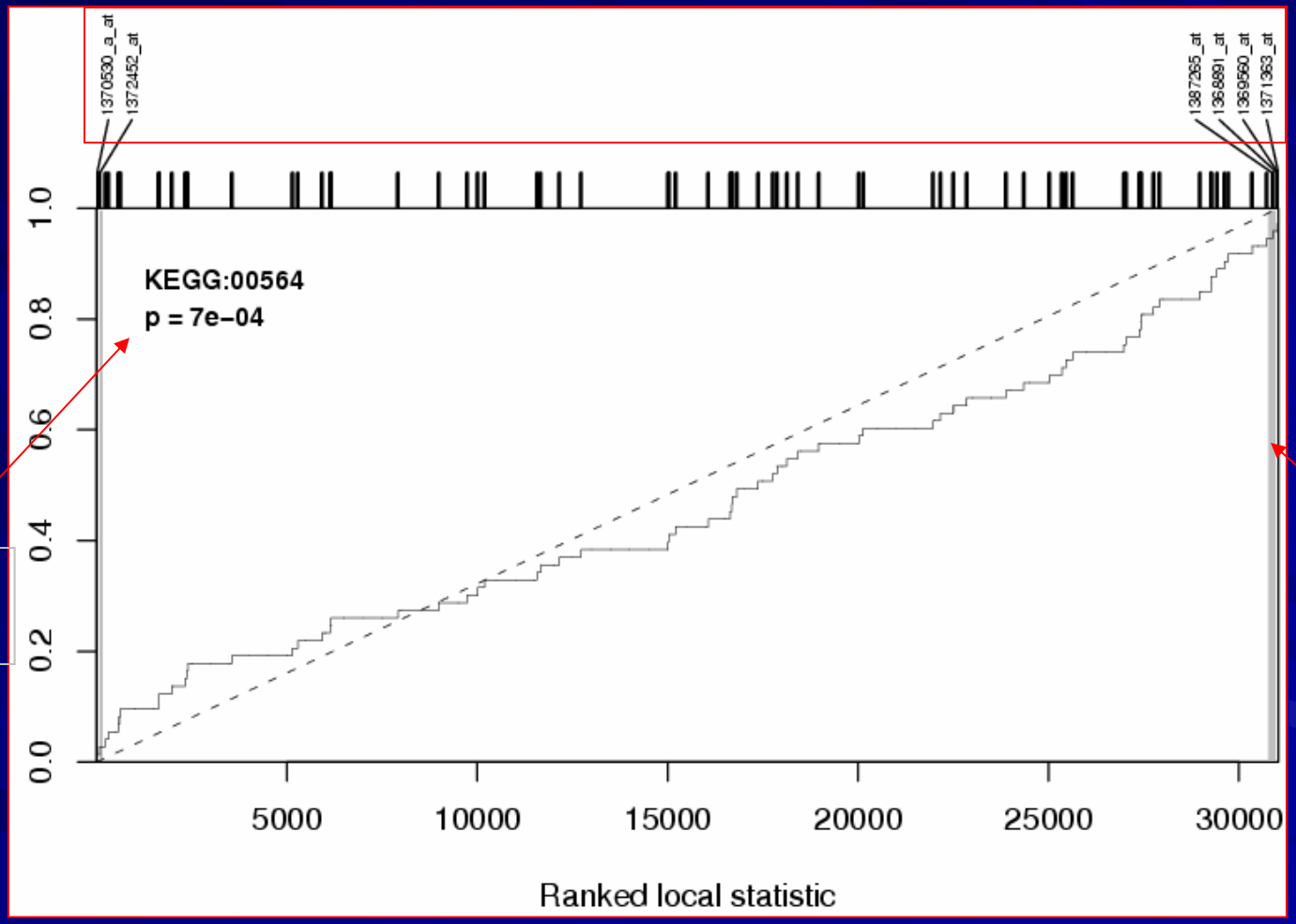
	Local.Stat	Emp.pvalue
1370530_a_at	-2.861	0.0083
1372452_at	-2.288	0.0283
1382986_at	-1.930	0.0592
1385209_at	-1.831	0.0753
1377398_at	-1.654	0.1039
1396648_at	-1.637	0.1051
1369758_at	-1.626	0.1105
.		
.		



SAFE Report: Detailed Pathway Information for Significant Category



SAFE Plot



Genes in Category

Category P-value

Shading indicates individually significant genes

Categories with Small p-values for Both Permethrin and Deltamethrin

GOBP

Category	Terms	DLTEmp.pvalue	PermEmp.pvalue
GO:0048754	branching morphogenesis of a tube	0.0171	2.00E-04
GO:0001763	morphogenesis of a branching structure	0.0172	2.00E-04
GO:0001569	patterning of blood vessels	0.0661	3.00E-04
GO:0007162	negative regulation of cell adhesion	0.0175	0.0025
GO:0009880	embryonic pattern specification	0.1259	5.00E-04
GO:0015718	monocarboxylic acid transport	0.0051	0.0125
GO:0007498	mesoderm development	0.0105	0.0067

GOCC

Category	Terms	DLTEmp.pvalue	PermEmp.pvalue
GO:0005954	calcium- and calmodulin-dependent protein kinase complex	0.0053	0.0146

GOMF

Category	Terms	DLTEmp.pvalue	PermEmp.pvalue
GO:0046915	transition metal ion transporter activity	0.0026	0.0348

PFAM:

Category	Terms	DLTEmp.pvalue	PermEmp.pvalue
PFAM:05210	Sprouty protein (Spry)	0.0401	0.0001
PFAM:03137	Organic Anion Transporter Polypeptide (OATP) family	0.0017	0.0087

A Peek at SAFE in ToxCast Data Analysis

Purpose of the Analysis:

To find significant pathway categories in treatment (dosed) vs control studies.

Data Source: ToxCast Data

Microarray: Affymetrix Rat Genome 230 2.0 Array
Affymetrix Human 133 Plus 2 Array

Control: Vehicle Controls

Treatment: propioconazole 100 uM (rat), propioconazole 100 uM (human), triadimefon 100 uM (rat), monoethylhexly pthalate 100 uM (rat).

Local Statistics: Limma moderated t–statistics

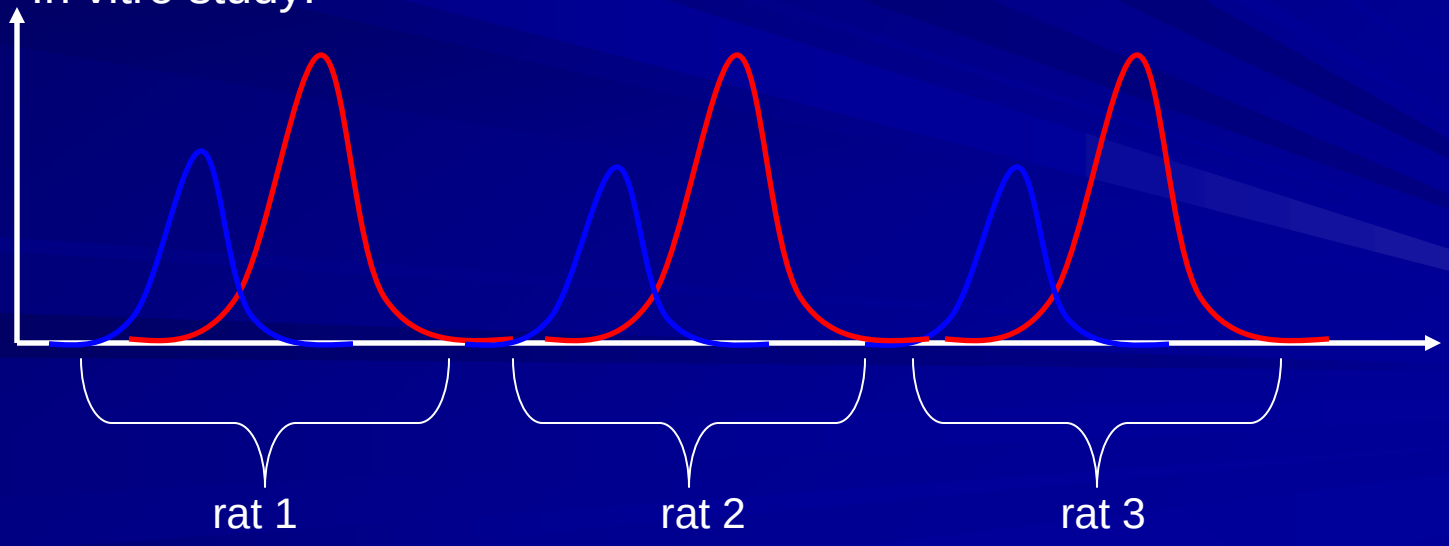
$$\tilde{t}_g = \frac{\overline{M}_g}{\tilde{s}_g / \sqrt{n}}$$

Why Limma moderated t-statistics?

To improve power in small sample sizes by borrowing information across genes.

Data Sample size: 3 arrays per condition for rat and 4 for human

In vitro study:



ToxCast SAFE Analysis

propioconazole 100 uM (rat)

Category	Adj.P-value	Name
KEGG00564	0.040036911	Glycerophospholipid metabolism
KEGG00190	0.040036911	Oxidative phosphorylation
KEGG00193	0.040036911	ATP synthesis
KEGG00440	0.040036911	Aminophosphonate metabolism
KEGG00521	0.040036911	Streptomycin biosynthesis
KEGG00601	0.040036911	Glycosphingolipid biosynthesis - lactoseries
KEGG03010	0.040036911	Ribosome
KEGG03030	0.040036911	DNA polymerase
KEGG00052	0.059045795	Galactose metabolism
KEGG00051	0.074835596	Fructose and mannose metabolism
KEGG00240	0.074835596	Pyrimidine metabolism
KEGG04710	0.083605756	Circadian rhythm

propioconazole 100 uM(rat)

Category	Adj.P-value	Name
KEGG00601	0.003687788	Glycosphingolipid biosynthesis - lactoseries
KEGG00970	0.003687788	Aminoacyl-tRNA biosynthesis
KEGG00623	0.012174554	2,4-Dichlorobenzoate degradation
KEGG04742	0.061066719	Taste transduction

ToxCast SAFE Analysis

triadimefon 100 uM (rat)

Category	Adj.P-value	Name
KEGG00100	0.00688486	Biosynthesis of steroids
KEGG00190	0.00688486	Oxidative phosphorylation
KEGG00532	0.00688486	Chondroitin sulfate biosynthesis
KEGG00533	0.00688486	Keratan sulfate biosynthesis
KEGG00970	0.00688486	Aminoacyl-tRNA biosynthesis
KEGG03010	0.00688486	Ribosome
KEGG03020	0.00688486	RNA polymerase
KEGG05060	0.00688486	Prion disease
KEGG01510	0.03117154	Neurodegenerative Disorders
KEGG00534	0.0698451	Heparan sulfate biosynthesis
KEGG00601	0.0698451	Glycosphingolipid biosynthesis - lactoseries
KEGG00531	0.08399337	Glycosaminoglycan degradation

monoethylhexly pthalate 100 uM (rat)

Category	Adj.P-value	Name
KEGG00020	0.04351446	Citrate cycle (TCA cycle)
KEGG03030	0.04351446	DNA polymerase

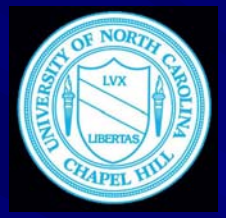
Acknowledgments

EPA

Richard Judson

Imran Shah

David Dix



UNC

Andrew Nobel

Mayetri Gupta

Ivan Rusyn

Daniel Gatti

DUKE

Bill Barry

Contact: Zhen Li
zli@bios.unc.edu

References

Barry WT, Nobel AB, and Wright FA (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**:1943-1949.

Beißbarth,T. and Speed,T.P. (2004) Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Bhattacharjee,A., Richards,W.G., Staunton,J., Li,C., Monti,S., Vasa,P., Ladd,C., Beheshti,J., Bueno,R., Gillette,M. *et al*. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl Acad. Sci. USA*, **98**, 13790–13795.

Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al*. (2003) Pgc-1alpharesponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Virtaneva,K.I., Wright,F.A., Tanner,S.M., Yuan,B., Lemon,W.J., Caligiuri,M.A., Bloomfield,C.D., de la Chapelle,A. and Krahe,R. (2001) Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics. *Proc. Natl Acad. Sci. USA*, **98**, 1124–1129.

Westfall,P.H. andYoung,S.S. (1989)*P*-value adjustment for multiple tests in multivariate binomial models. *J. Amer. Statist. Assoc.*, **84**, 780–786.

Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery

References, cont.

rate controlling multiple test procedures for correlated test statistics. *J. Statist. Plann. Inference*, **82**, 171–196.

Zeeberg,B.R., Feng,W.,Wang,G.,Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al*. (2003) Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.

Zhong,S., Tian,L., Li,C., Storch,F.K. and Wong,W.H. (2004) Comparative analysis of gene sets in the gene ontology space under the multiple hypothesis testing framework. *Proc. IEEE Comput. Syst. Bioinformatics*, in press.