Pathway and Gene Set Analysis Part 2

Alison Motsinger-Reif, PhD
Bioinformatics Research Center
Department of Statistics
North Carolina State University
motsinger@stat.ncsu.edu

Goals

Some methods in more detail

- TopGO
- Global Ancova
- Pathvisio/Genmapp
- Impact Factor Analysis
- GSEA

Some methods in detail

- There are far too many methods to give a comprehensive overview

BRIEFINGS IN BIOINFORMATICS. VOL 9. NO 3. 189–197
Advance Access publication January 17, 2008
doi:10.1093/bib/bbn001

Gene-set approach for expression pattern analysis

Dougu Nam and Seon-Young Kim

Submitted: 7th November 2007; Received (in revised form): 28th December 2007

Abstract

Recently developed gene set analysis methods evaluate differential expression patterns of gene groups instead of those of individual genes. This approach especially targets gene groups whose constituents show subtle but coordinated expression changes, which might not be detected by the usual individual gene analysis. The approach has been quite successful in deriving new information from expression data, and a number of methods and tools have been developed intensively in recent years. We review those methods and currently available tools, classify them according to the statistical methods employed, and discuss their pros and cons. We also discuss several interesting extensions to the methods.

Keywords: *gene set analysis; DNA microarray; differential expression of genes*

Table of methods (from Nam & Kim)

Table I: Cutoff-free gene set analysis methods

Authors	Year	Name	Statistical test	Self-contained versus competitive	Gene versus ample randomization	Reference
Virtaneva *et al.*	2001		sample randomization	self-contained	sample	[8]
Pavlidis *et al.*	2002		gene randomization	competitive	gene	[9]
Mootha *et al.*	2003	GSEA	sample randomization	mixed	sample	[7]
Breslin *et al.*	2004	Catmap	gene randomization	competitive	gene	[3]
Goeman *et al.*	2004	globaltest	sample randomization	self-contained	sample	[17]
Smid *et al.*	2004	GO-Mapper	z-test	competitive	gene	[38]
Volinia *et al.*	2004	GOAL	gene randomization	competitive	gene	[39]
Barry *et al.*	2005	SAFE	sample randomization	competitive	sample	[19]
Beh-Shaul *et al.*	2005		Kolmogorov–Smirnov test	competitive	gene	[5]
Boorsma e al.	2005	T-profiler	*t*-test	competitive	gene	[15]
Kim *et al.*	2005	PAGE	z-test	competitive	gene	[14]
Lee *et al.*	2005	ErmineJ	sample randomization	competitive	gene	[16]
Subramanian *et al.*	2005	GSEA	sample randomization	mixed	gene	[25]
Tian *et al.*	2005	Q1, Q2	gene or sample randomization	competitive or self-contained	gene or sample	[10]
Tomfohr *et al.*	2005	PLAGE	sample randomization	self-contained	sample	[20]
Edelman *et al.*	2006	ASSESS	sample randomization	competitive	sample	[28]
Kong *et al.*	2006		Hotelling's T squared	self-contained	sample	[21]
Nam *et al.*	2006	ADGO	z-test	competitive	gene	[29]
Saxena *et al.*	2006	AE	sample randomization	competitive	sample	[31]
Scheer *et al.*	2006	JProGO	Fisher's exact test, Kolmogorov–Smirnov test, *t*-test, unpaired Wilcoxon's test	competitive	gene	[40]
Al-Shahrour *et al.*	2007	Fatiscan	Fisher's exact test, hypergeometric test	competitive	gene	[41]
Backes *et al.*	2007	GeneTrail	Fisher's exact test, hypergeometric test, sample randomization	competitive	gene or sample	[42]
Cavalieri *et al.*	2007	Eu.Gene Analyzer	Fisher's exact test, sample randomization	competitive	gene or sample	[43]
Dinu *et al.*	2007	SAM-GS	sample randomization	self-contained	sample	[22]
Efron *et al.*	2007	GSA	sample randomization	mixed	sample	[26]
Newton *et al.*	2007	Random set	z-test	competitive	gene	[44]

Table of software (from Nam & Kim)

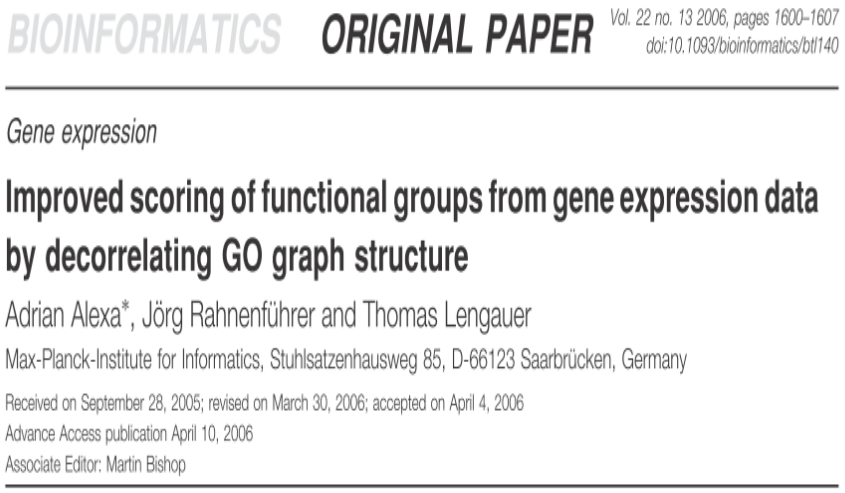
Table 2: Gene set analysis tools

Name	Organism[a]	Application Type	URL	Reference
ADGO	H, M, R, Y	Web server	http://array.kobic.re.kr/ADGO	[29]
ASSESS	H, M, R	Octave/Java standalone	http://people.genome.duke.edu/~jhg9/assess/	[28]
Babelomics	H, M, R, DM, S, C	Web server	http://www.babelomics.org	[45]
Catmap	H	Perl script	http://bioinfo.thep.lu.se/catmap.html	[3]
ErmineJ	H, M, R	Java standalone	http://www.bioinformatics.ubc.ca/ermineJ/	[16]
Eu.Gene Analyzer	H, M, R, Y	Windows/Unix standalone	http://www.ducciocavalieri.org/bio/Eugene.htm	[43]
FatiScan	H, M, R, Y, B, D, G, C, A, S, DM	Web server	http://fatiscan.bioinfo.cipf.es/	[41]
GAZER	H, M, R, Y	Web server	http://integromics.kobic.re.kr/GAzer/index.faces;	[13]
GeneTrail	H, M, R, Y, SA, CG, AT	Web server	http://genetrail.bioinf.uni-sb.de/	[42]
Global test	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/globaltest.html	[17]
GOAL	H, M	Web server	http://microarrays.unife.it	[39]
GO-Mapper	H, M, R, Z, DM, Y	Windows standalone, Perl script	http://www.gatcplatform.nl/	[38]
GSA	H	R package	http://www-stat.stanford.edu/~tibs/GSA/	[26]
GSEA	H	Java standalone, R package	http://www.broad.mit.edu/gsea/	[25]
JProGO	Various prokaryotes	Web server	http://www.jprogo.de/	[40]
MEGO	H	Windows standalone	http://www.dxy.cn/mego/	[46]
PAGE	H, M, R, Y	Python script	From the author (kimsy@kribb.re.kr)	[14]
PLAGE	H, M	Web server	http://dulci.biostat.duke.edu/pathways/	[20]
SAFE	NA	R package	http://bioconductor.org/packages/2.0/bioc/html/safe.html	[19]
SAM-GS	NA	Windows Excel Add-In	http://www.ualberta.ca/~yyasui/homepage.html	[22]
T-profiler	Y, CA	Web server	http://www.t-profiler.org/	[15]

[a]H: *Homo sapiens*; M: *Mus musculus*; R: *Rattus norvegicus*; Y: *Sacchaomyces cerevisiae*; B: *Bos Taurus*; D: *Daniel rerio*; G: *Gallus gallus*; C: *Caenorhabditis elegans*; A: *Arabidopsis thaliana*; DM: *Drosophila melanogaster*; Z: *Zebra fish*; CA: *Candida albicans*; SA: *Staphylococcus aureus*; CG: *Corynebacterium glutamicum*; AT: *Arabidopsis thaliana*.

TopGO

- TopGO is a GO term analysis program available from Bioconductor
- It takes the GO hierarchy into account when scoring terms
- If a parent term is only significant because of child term, it will receive a lower score
- TopGO uses the Fisher-test or the KS-test (both competitive)
- TopGO also gives a graphical representation of the results in form of a tree



BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

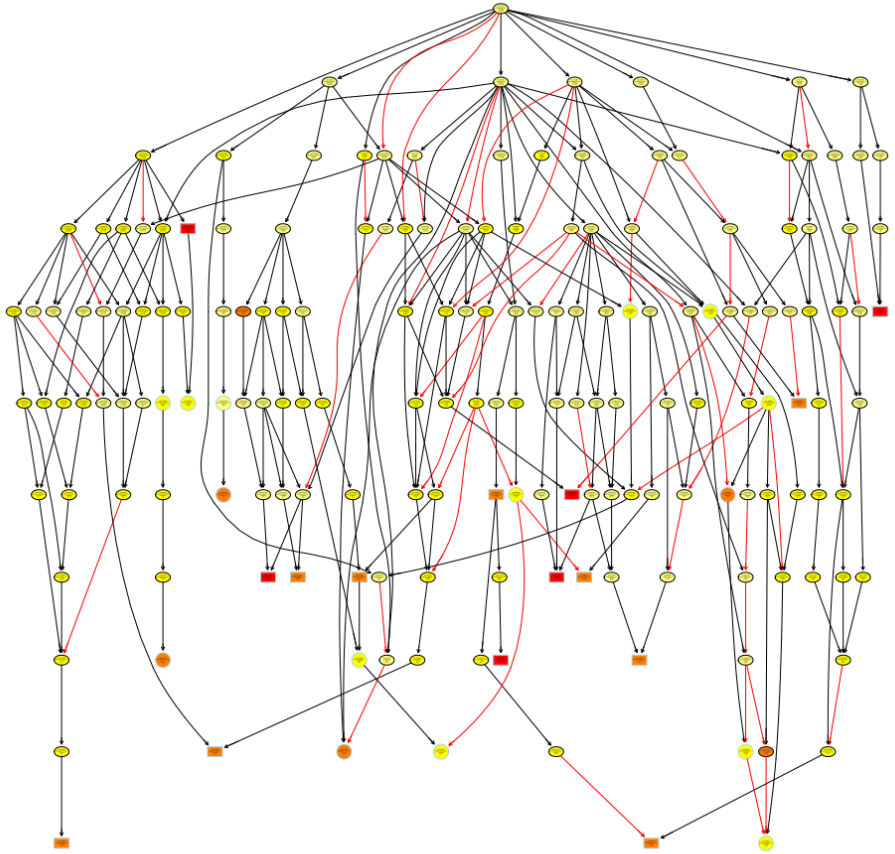
Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

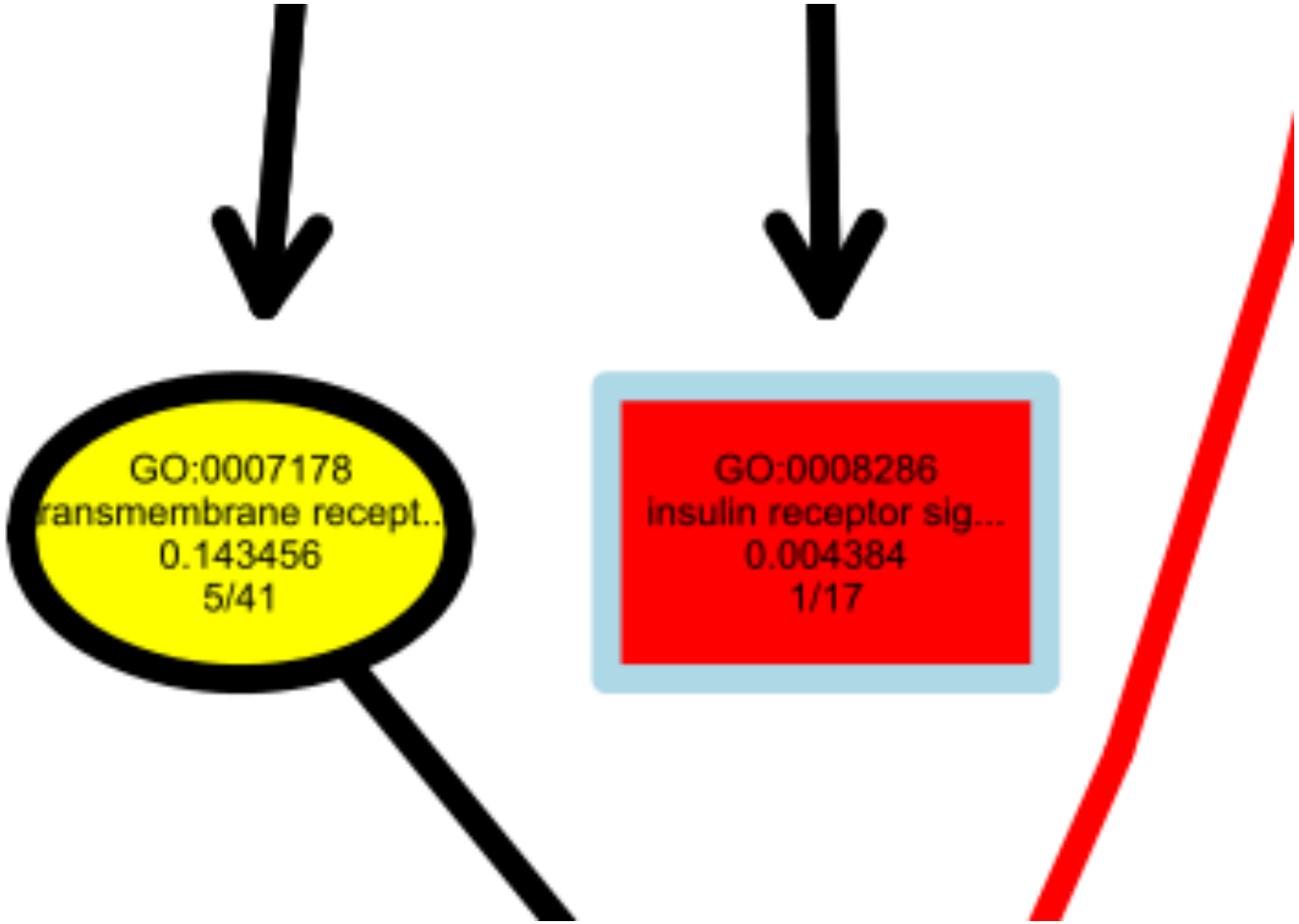
Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer
Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006
Advance Access publication April 10, 2006
Associate Editor: Martin Bishop

Tree showing the 15 most significant GO terms

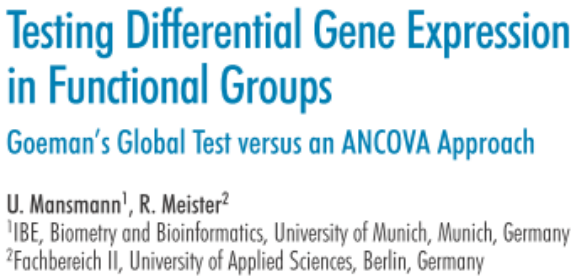


Zooming in


GO:0007178
ransmembrane recept..
0.143456
5/41
GO:0008286
insulin receptor sig...
0.004384
1/17

Global Ancova

- Uses all data (instead of summary statistics)
- NOT a multivariate method (MANOVA)
- One linear model for all genes within the gene set
  - Gene is a factor in the model that interacts with other factors
- Full model (e.g. including difference between lean and obsese) is compared with restricted model (no difference)
- P-values are calculated by group label resampling
- Algorithm allows for complex linear models including covariates
- Related to Goeman's Globaltest, which reverses roles of gene expression and groups: Goeman uses gene expression to explain groups (logistic regression)


Testing Differential Gene Expression in Functional Groups

Goeman's Global Test versus an ANCOVA Approach

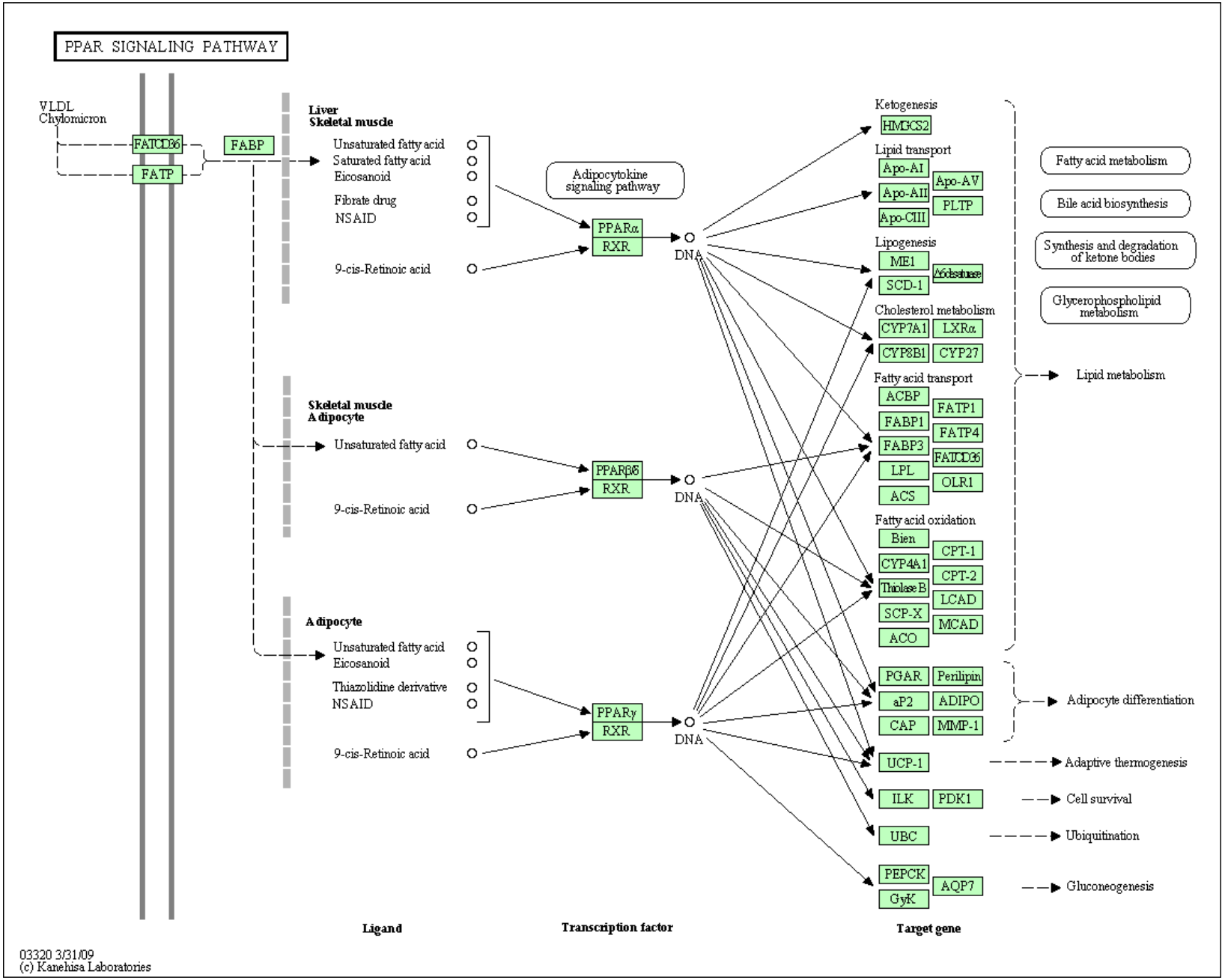
U. Mansmann[1], R. Meister[2]
[1]IBE, Biometry and Bioinformatics, University of Munich, Munich, Germany
[2]Fachbereich II, University of Applied Sciences, Berlin, Germany

10 most significant KEGG pathways according to Global Ancova

Pathway Name	path.size	sig.genes	perc.sig	p.gs	p.fisher	p.globaltest	p.globalAncova
Pantothenate and CoA biosynthesis	11	3	27.27%	7.05%	9.08%	0.55%	0.01%
Valine, leucine and isoleucine biosynthesis	4	2	50.00%	4.10%	5.29%	0.22%	0.02%
Cell Communication	60	10	16.67%	8.77%	7.51%	1.02%	0.03%
PPAR signaling pathway	37	10	27.03%	11.01%	0.28%	1.64%	0.07%
Inositol metabolism	1	1	100.00%	8.46%	10.06%	0.19%	0.10%
Valine, leucine and isoleucine degradation	35	7	20.00%	49.56%	5.65%	1.42%	0.11%
Fatty acid metabolism	27	6	22.22%	49.59%	4.81%	1.54%	0.31%
ECM-receptor interaction	49	8	16.33%	4.91%	11.45%	1.47%	0.83%
Focal adhesion	122	16	13.11%	76.63%	16.40%	2.59%	0.87%
Purine metabolism	78	14	17.95%	26.82%	2.26%	3.42%	1.21%

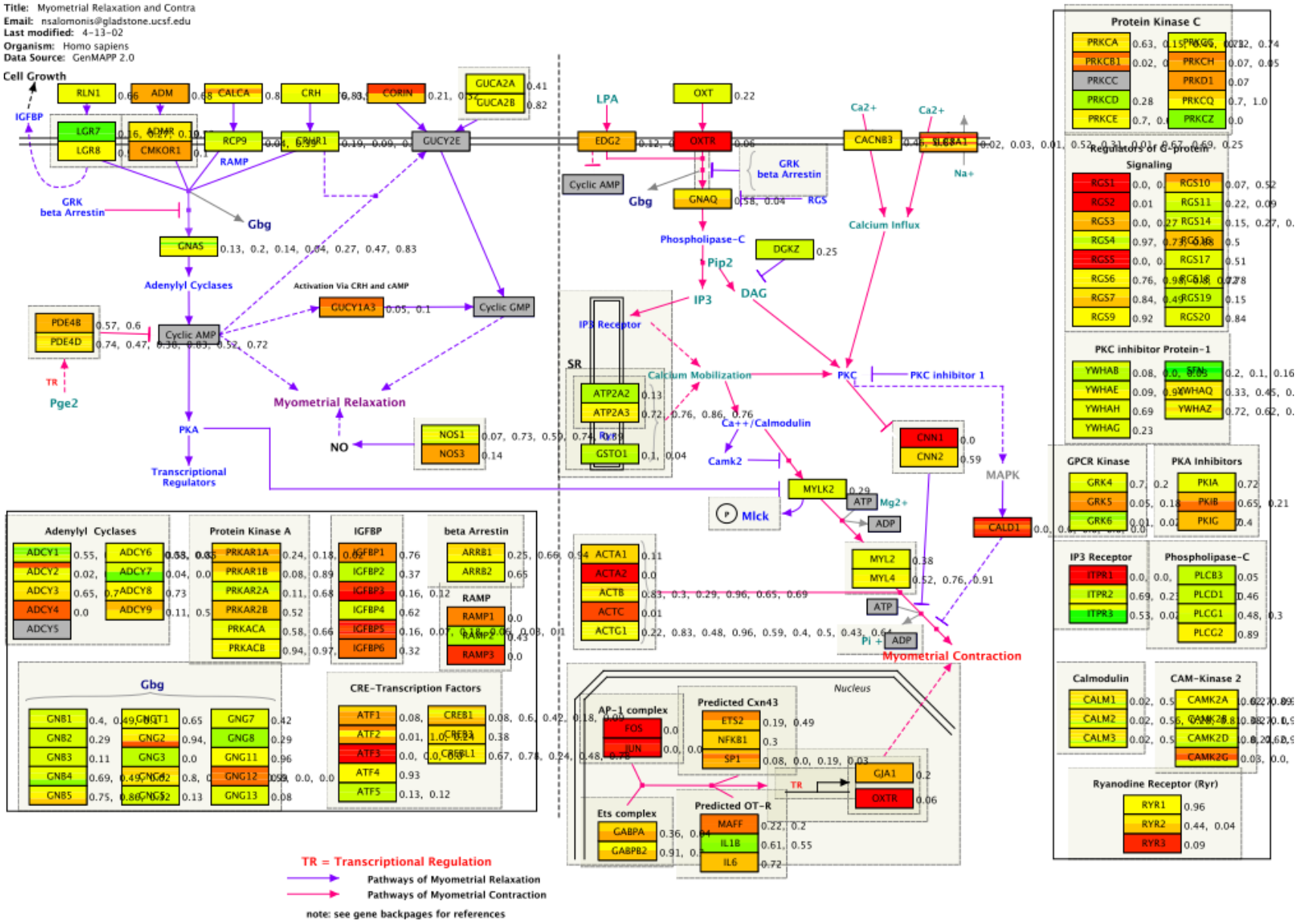
p.gs = A GSEA related competitive method (available in Limma)

p.fisher = Fisher-Test (competitive)


PPAR SIGNALING PATHWAY
VLDL
Chylomicron
FATCD36
FATP
FABP
Liver
Skeletal muscle
Unsaturated fatty acid
Saturated fatty acid
Eicosanoid
Fibrate drug
NSAID
9-cis-Retinoic acid
Adipocytokine signaling pathway
PPARα
RXR
DNA
Skeletal muscle
Adipocyte
Unsaturated fatty acid
9-cis-Retinoic acid
PPARβ/δ
RXR
DNA
Adipocyte
Unsaturated fatty acid
Eicosanoid
Thiazolidine derivative
NSAID
9-cis-Retinoic acid
PPARγ
RXR
DNA
Ketogenesis
HMGCS2
Lipid transport
Apo-AI
Apo-AII
Apo-CIII
Apo-AV
PLTP
Lipogenesis
ME1
SCD-1
Acetyl-CoA carboxylase
Cholesterol metabolism
CYP7A1
LXRα
CYP8B1
CYP27
Fatty acid transport
ACBP
FABP1
FABP3
LPL
ACS
FATP1
FATP4
FATCD36
OLR1
Fatty acid oxidation
Bien
CYP4A1
Thiolase B
SCP-X
ACO
CPT-1
CPT-2
LCAD
MCAD
PGAR
aP2
CAP
Perilipin
ADIPO
MMP-1
UCP-1
ILK
PDK1
UBC
PEPCK
GyK
AQP7
Fatty acid metabolism
Bile acid biosynthesis
Synthesis and degradation of ketone bodies
Glycerophospholipid metabolism
Lipid metabolism
Adipocyte differentiation
Adaptive thermogenesis
Cell survival
Ubiquitination
Gluconeogenesis
Ligand
Transcription factor
Target gene
03320 3/31/09
(c) Kanehisa Laboratories

Genmapp/Pathvisio

- These are two pathway visualisation tools that collaborate
  - http://www.genmapp.org
  - http://www.pathvisio.org
- Both do some basic statistical analysis too (Fisher-Test with normal approximation)
- Main focus is on visually displaying pathways
  - Genes/nodes can be color-coded according to the data
  - Results (p-values, fold changes) can be displayed next to genes/nodes


Title: Myometrial Relaxation and Contra
Email: nsalomonis@gladstone.ucsf.edu
Last modified: 4-13-02
Organism: Homo sapiens
Data Source: GenMAPP 2.0
Cell Growth
IGFBP
RLN1
ADM
CALCA
CRH
CORIN
GUCA2A
GUCA2B
LGR7
LGR8
ADMR
CMKOR1
RCP9
CNR1
GUCY2E
RAMP
GRK
beta Arrestin
Gbg
GNAS
Adenylyl Cyclases
Activation Via CRH and cAMP
GUCY1A3
Cyclic GMP
PDE4B
PDE4D
Cyclic AMP
TR
Pge2
Myometrial Relaxation
PKA
NO
NOS1
NOS3
Transcriptional
Regulators
LPA
OXT
EDG2
OXTR
GRK
beta Arrestin
RGS
Cyclic AMP
Gbg
GNAQ
Ca2+
Ca2+
CACNB3
SLC8A1
Na+
Calcium Influx
Phospholipase-C
DGKZ
Pip2
DAG
IP3
IP3 Receptor
SR
ATP2A2
ATP2A3
GSTO1
Calcium Mobilization
PKC
PKC inhibitor 1
Ca++/Calmodulin
Camk2
MYLK2
ATP
Mg2+
ADP
Mlck
CNN1
CNN2
MAPK
CALD1
ACTA1
ACTA2
ACTB
ACTC
ACTG1
MYL2
MYL4
ATP
ADP
Pi +
Myometrial Contraction
Nucleus
AP-1 complex
FOS
JUN
Predicted Cxn43
ETS2
NFKB1
SP1
TR
GJA1
OXTR
Predicted OT-R
MAFF
IL1B
IL6
Ets complex
GABPA
GABPB2
Protein Kinase C
PRKCA
PRKCB1
PRKCC
PRKCD
PRKCE
PRKCG
PRKCH
PRKD1
PRKCQ
PRKCZ
Regulators of G-protein
Signaling
RGS1
RGS2
RGS3
RGS4
RGS5
RGS6
RGS7
RGS9
RGS10
RGS11
RGS14
RGS16
RGS17
RGS18
RGS19
RGS20
PKC inhibitor Protein-1
YWHAB
YWHAE
YWHAH
YWHAG
SFN
YWHAQ
YWHAZ
GPCR Kinase
GRK4
GRK5
GRK6
PKA Inhibitors
PKIA
PKIB
PKIG
IP3 Receptor
ITPR1
ITPR2
ITPR3
Phospholipase-C
PLCB3
PLCD1
PLCG1
PLCG2
Calmodulin
CALM1
CALM2
CALM3
CAM-Kinase 2
CAMK2A
CAMK2B
CAMK2D
CAMK2G
Ryanodine Receptor (Ryr)
RYR1
RYR2
RYR3
Adenylyl Cyclases
ADCY1
ADCY2
ADCY3
ADCY4
ADCY5
ADCY6
ADCY7
ADCY8
ADCY9
Protein Kinase A
PRKAR1A
PRKAR1B
PRKAR2A
PRKAR2B
PRKACA
PRKACB
IGFBP
IGFBP1
IGFBP2
IGFBP3
IGFBP4
IGFBP5
IGFBP6
beta Arrestin
ARRB1
ARRB2
RAMP
RAMP1
RAMP2
RAMP3
Gbg
GNB1
GNB2
GNB3
GNB4
GNB5
GNGT1
GNG2
GNG3
GNG4
GNG5
GNG7
GNG8
GNG11
GNG12
GNG13
CRE-Transcription Factors
ATF1
ATF2
ATF3
ATF4
ATF5
CREB1
CREB3
CREBL1
TR = Transcriptional Regulation
Pathways of Myometrial Relaxation
Pathways of Myometrial Contraction
note: see gene backpages for references

Impact Factor Analysis

- Impact Factor (IF) analysis combines both ORA and FCS approach, while accounting for the topology of the pathway

- IF analysis computes Perturbation Factor (PF) for each gene in each pathway, which is a gene-level statistic, as follows:

$$PF(g_i) = \Delta F(g_i) + \sum_{j=1}^{n} \beta_{ji} \cdot \frac{PF(g_j)}{N_{ds}(g_j)}$$

- The first term, $\Delta F(g_i)$, represents the signed normalized measured expression change (i.e., fold change) of the gene g_i
- The second term accounts for the topology of the pathway, where gene g_j is upstream of gene g_i
- In the second term, β_{ji} represents the type and strength of interaction betweer g_j and g_i
- If g_j activates g_i, $\beta_{ji} = 1$, and if g_j inhibits g_i, $\beta_{ji} = -1$
- Note that the PF of the upstream gene g_j is normalized by the number of downstream genes it interacts with, $N_{ds}(g_i)$
- The second term is repeated for every gene g_j that is upstream of gene g_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 1st term captures the significance of the given pathway P_i as provided by ORA, where p_i corresponds to the probability of obtaining a value of the statistic used at least as extreme as the one observed when the null hypothesis is true

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Because IF should be large for severely impacted pathways (i.e., small p-values), the 1st term uses $1/p_i$ rather than p_i

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

Log function is necessary to map the exponential scale of the p-values to a linear scale in order to keep the model linear

Impact Factor Analysis

- Next, Impact Factor (IF), is computed:

$$IF(P_i) = log\left(\frac{1}{p_i}\right) + \frac{\left|\sum_{g \in P_i} PF(g)\right|}{N_{de}(P_i)}$$

The 2nd term sums up the values of the PFs for all genes *g* on the given pathway P_i, and is normalized by the number of differentially expressed genes on the given pathway P_i

Impact Factor Analysis

- Note that Eq. 1 essentially describes the perturbation factor PF for a gene g_i as a linear function of the perturbation factors of all genes in a given pathway
- Therefore, the set of all equations defining the PFs for all genes in a given pathway P_i form a system of simultaneous equations
- Expanding and re-arranging Equation 1 for all genes g_1, g_2, …., g_n in a pathway Pi can be re-written as follows:

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1-\frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1-\frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1-\frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1)\cdot\Delta E(g_1) \\ \alpha(g_2)\cdot\Delta E(g_2) \\ \cdots \\ \alpha(g_n)\cdot\Delta E(g_n) \end{pmatrix}$$

Impact Factor Analysis

$$\begin{pmatrix} PF(g_1) \\ PF(g_2) \\ \cdots \\ PF(g_n) \end{pmatrix} = \begin{pmatrix} 1-\frac{\beta_{11}}{N_{ds(g_1)}} & -\frac{\beta_{21}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n1}}{N_{ds(g_n)}} \\ -\frac{\beta_{12}}{N_{ds(g_1)}} & 1-\frac{\beta_{22}}{N_{ds(g_2)}} & \cdots & -\frac{\beta_{n2}}{N_{ds(g_n)}} \\ \cdots & \cdots & \cdots & \cdots \\ -\frac{\beta_{1n}}{N_{ds(g_1)}} & -\frac{\beta_{2n}}{N_{ds(g_2)}} & \cdots & 1-\frac{\beta_{nn}}{N_{ds(g_n)}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha(g_1)\cdot\Delta E(g_1) \\ \alpha(g_2)\cdot\Delta E(g_2) \\ \cdots \\ \alpha(g_n)\cdot\Delta E(g_n) \end{pmatrix}$$

- After computing the PFs of all genes in a given pathway as the solution of this linear system, Eq. 2 is used to calculate the impact factor of each pathway

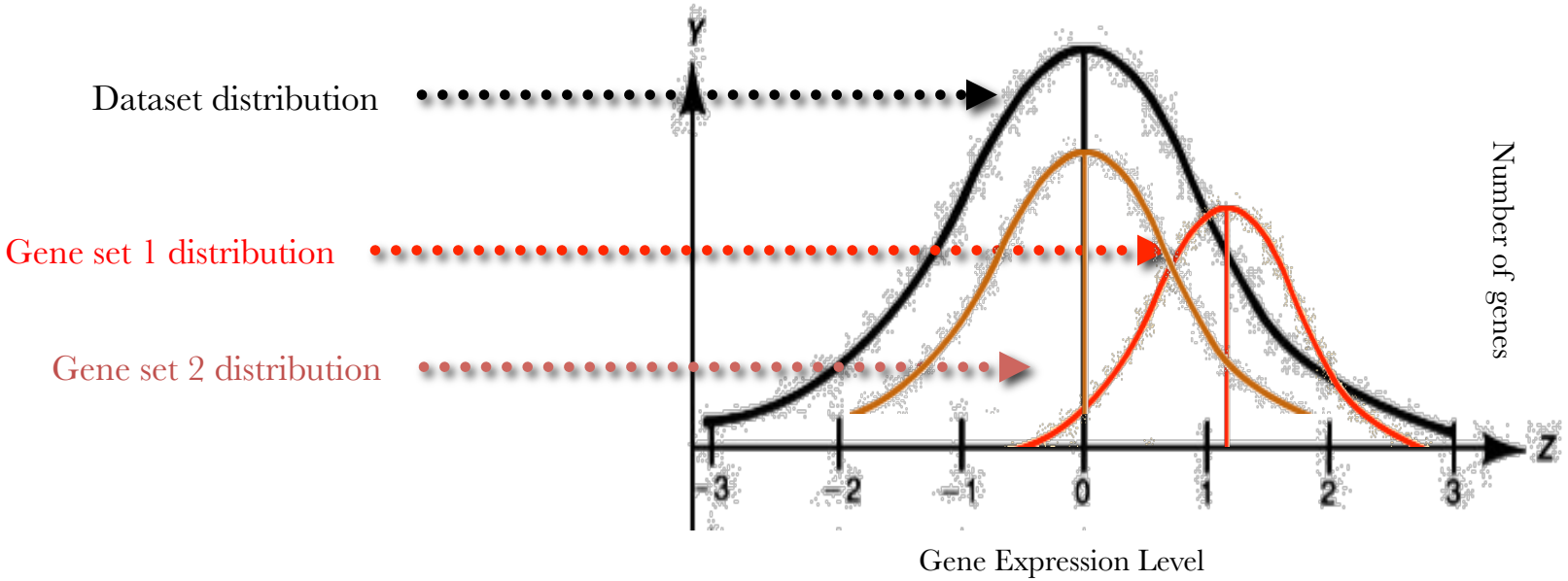
- The impact factor of each pathway is then used as a score to assess the impact of a given gene expression data set on all pathways (the higher the impact factor the more significant the pathway)

Gene Set Enrichment Analysis (GSEA)

- GSEA can be used with any gene set
- It is available as a standalone program, and versions of GSEA available within R/Bioconductor
- GSEA has many options and is a mix of a competitive and self-contained method
  - Default methods is to use a Kolmogorov Smirnov-type statistic to test the distribution of the gene set in the ranked gene list (competitive)
  - Typically that statistic ("enrichment score") is tested by permuting/ reshuffling the group labels (self-contained)
- Two Key Papers
  - Mootha et al., Nature Genetics 34, 267–273 (2003)
  - Subramanian et al., PNAS 102(43), 15545–15550 (2005).
    - Note - the description of GSEA changed between the two papers.

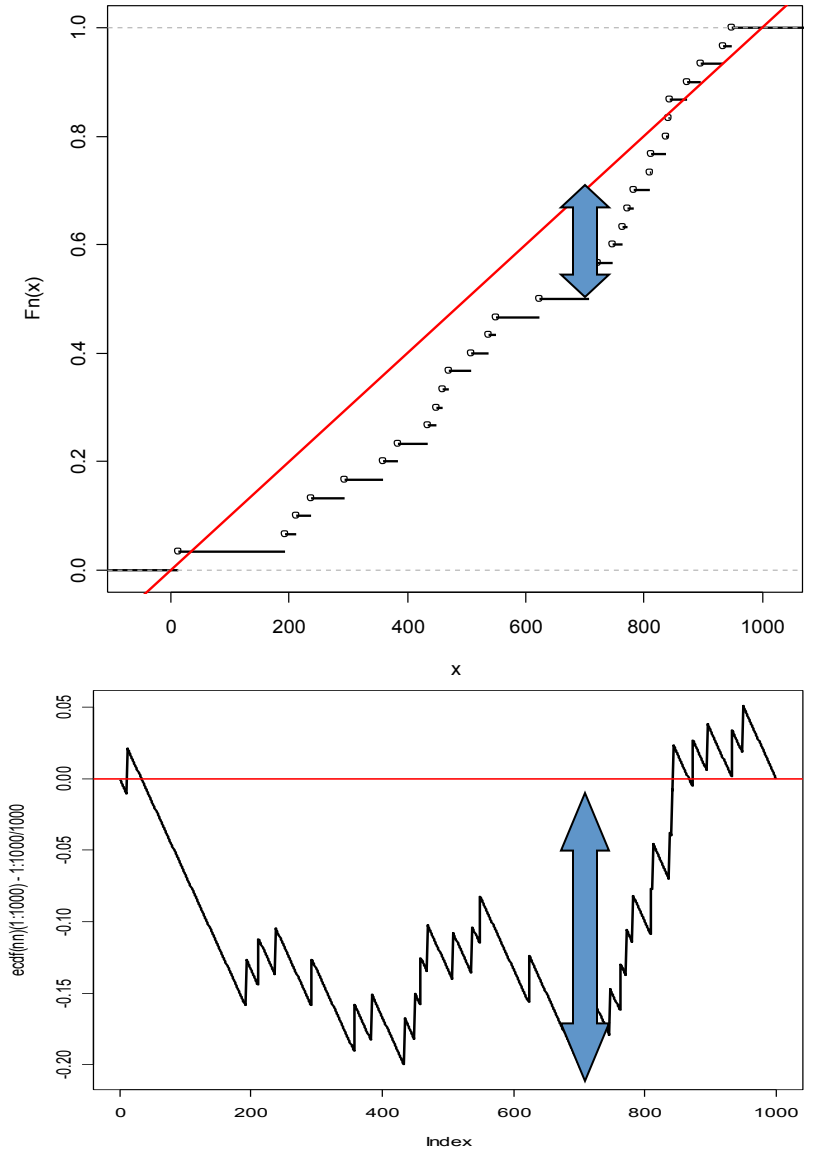
K-S Test

The Kolmogorov–Smirnov test is used to determine whether two underlying one-dimensional probability distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution, in either case based on finite samples.

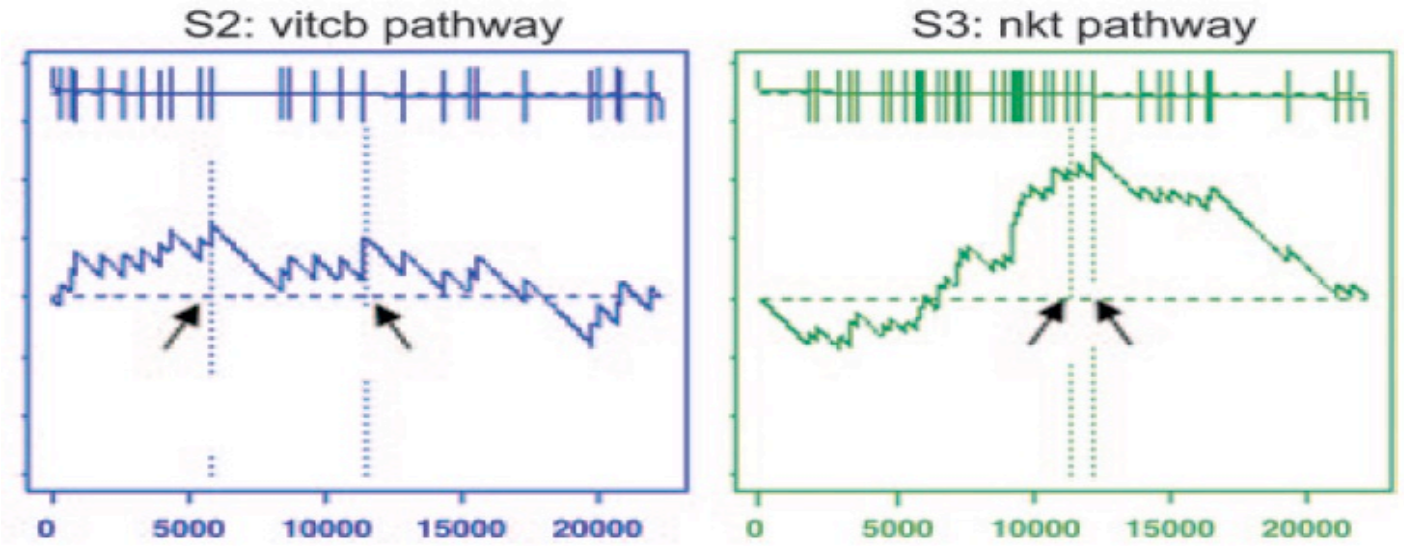


Kolmogorov-Smirnov Test

- Based on statistics of ‘Brownian Bridge’
  - random walk fixed end
- Maximum difference is test statistic
  - Null distribution known
- Reformulated by GSEA as difference of CDF – uniform from axis



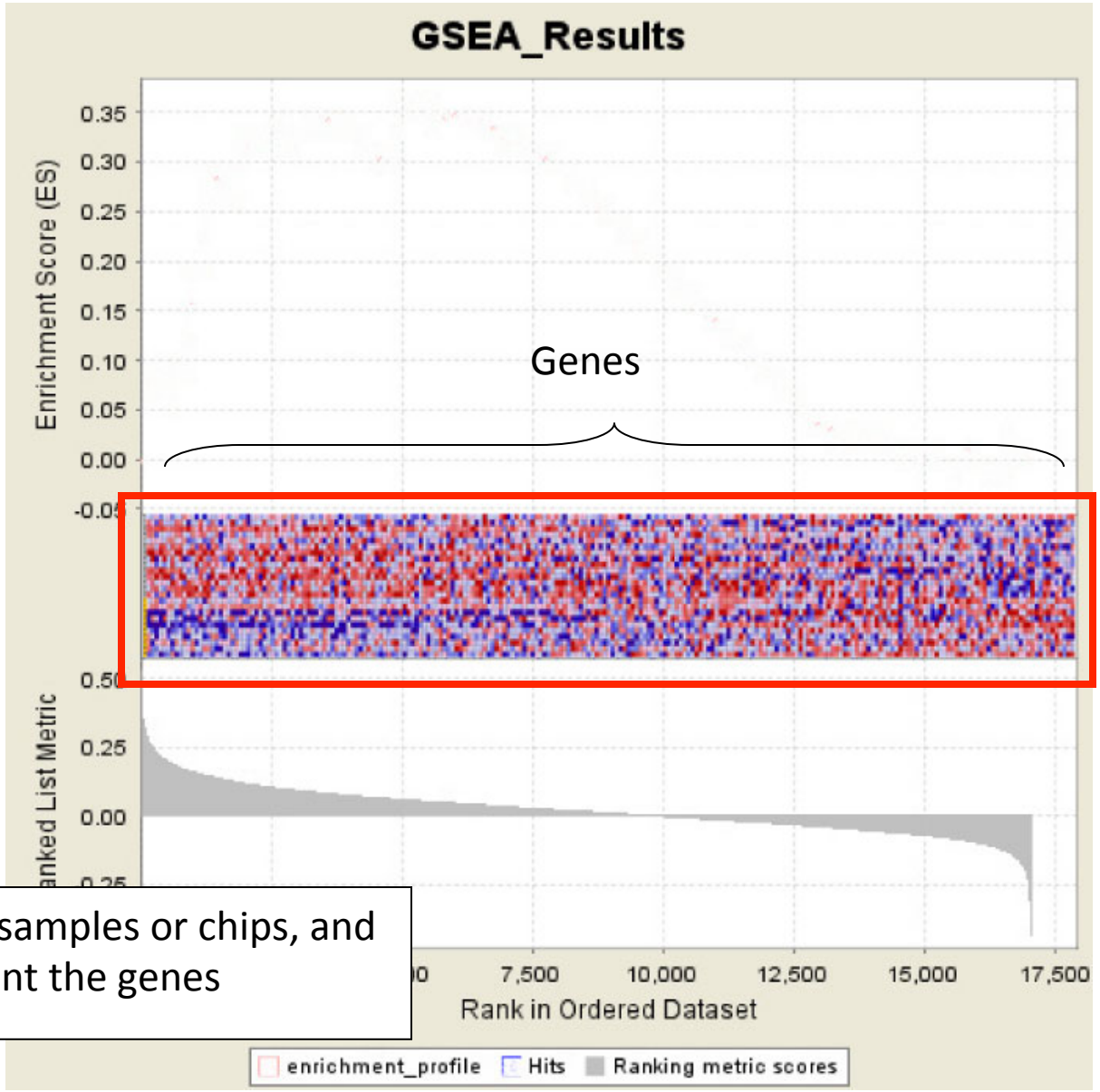
K-S Test Finds Irrelevant Sets



- Sometimes ranks concentrated in middle
  - K-S statistic high, but not meaningful for path change
- Fix: ad-hoc weighting by actual t-scores emphasizes departures at extreme ends
- No theory
- Generate null distribution by permutation

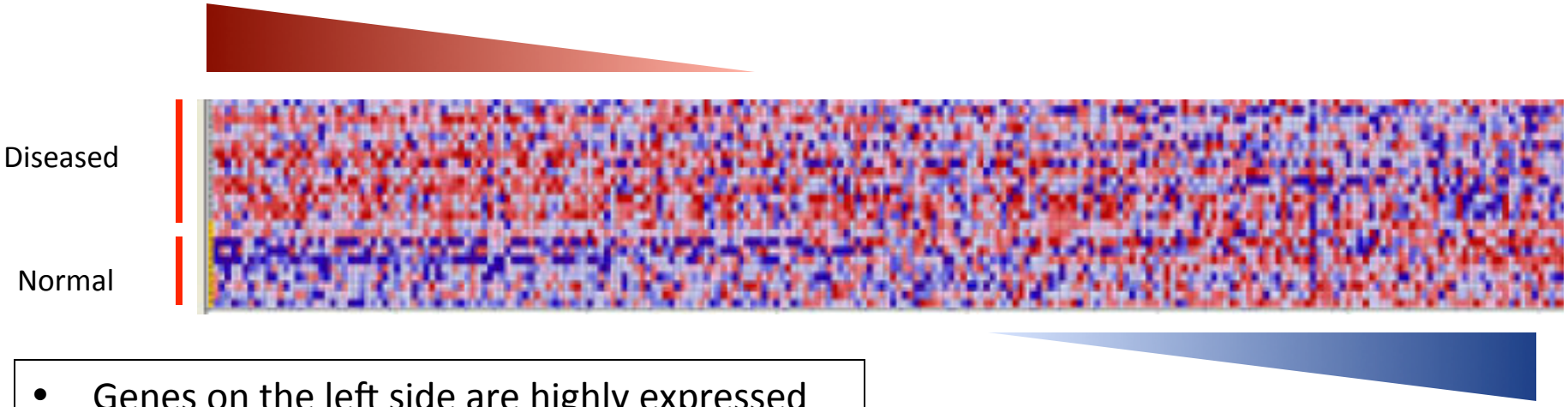
GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype.
- Record the maximum deviation from zero as the enrichment score



The rows represent the samples or chips, and the columns represent the genes

Highly expressed in diseased



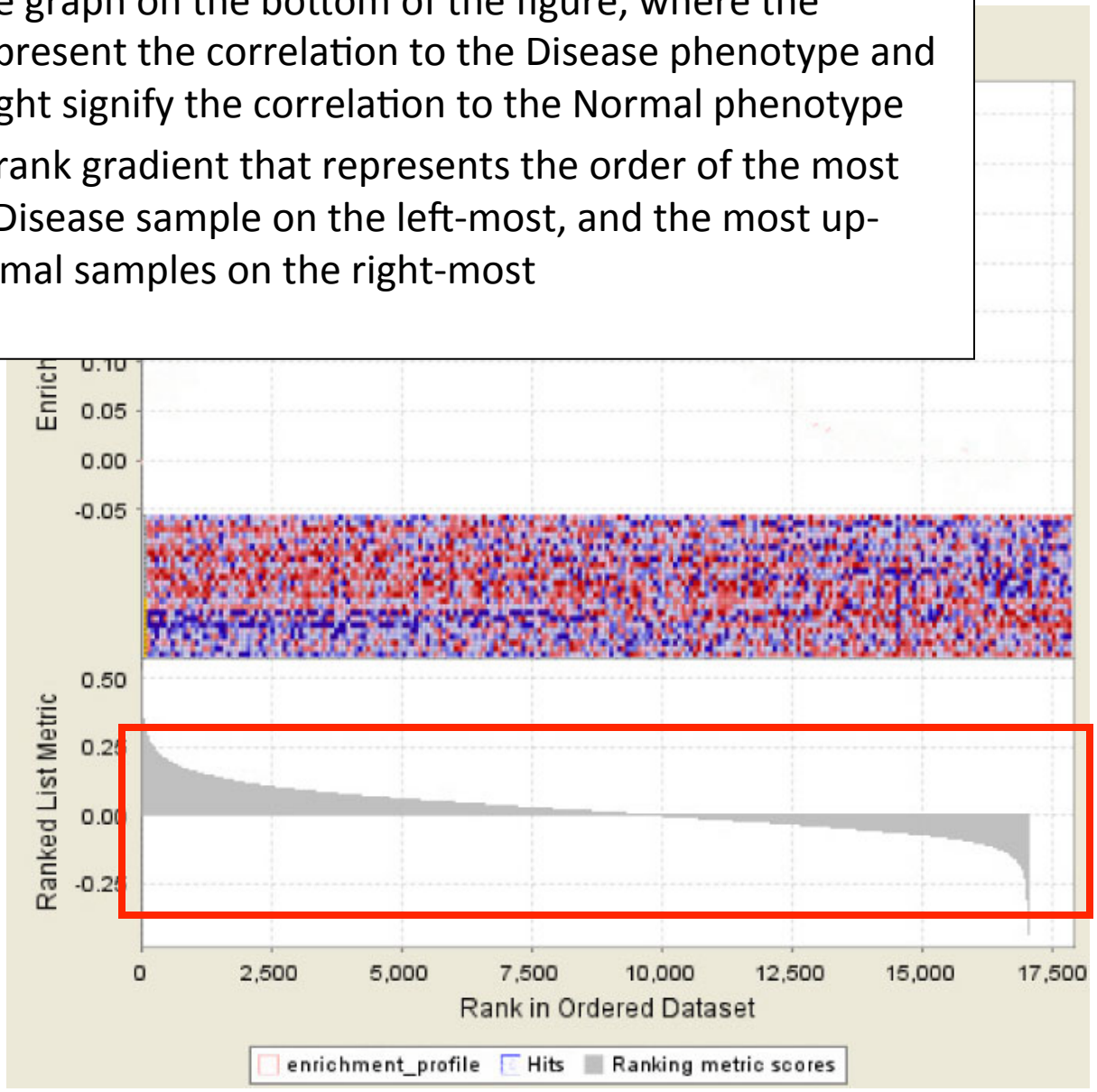
Lowly expressed in diseased

- Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes
- Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes

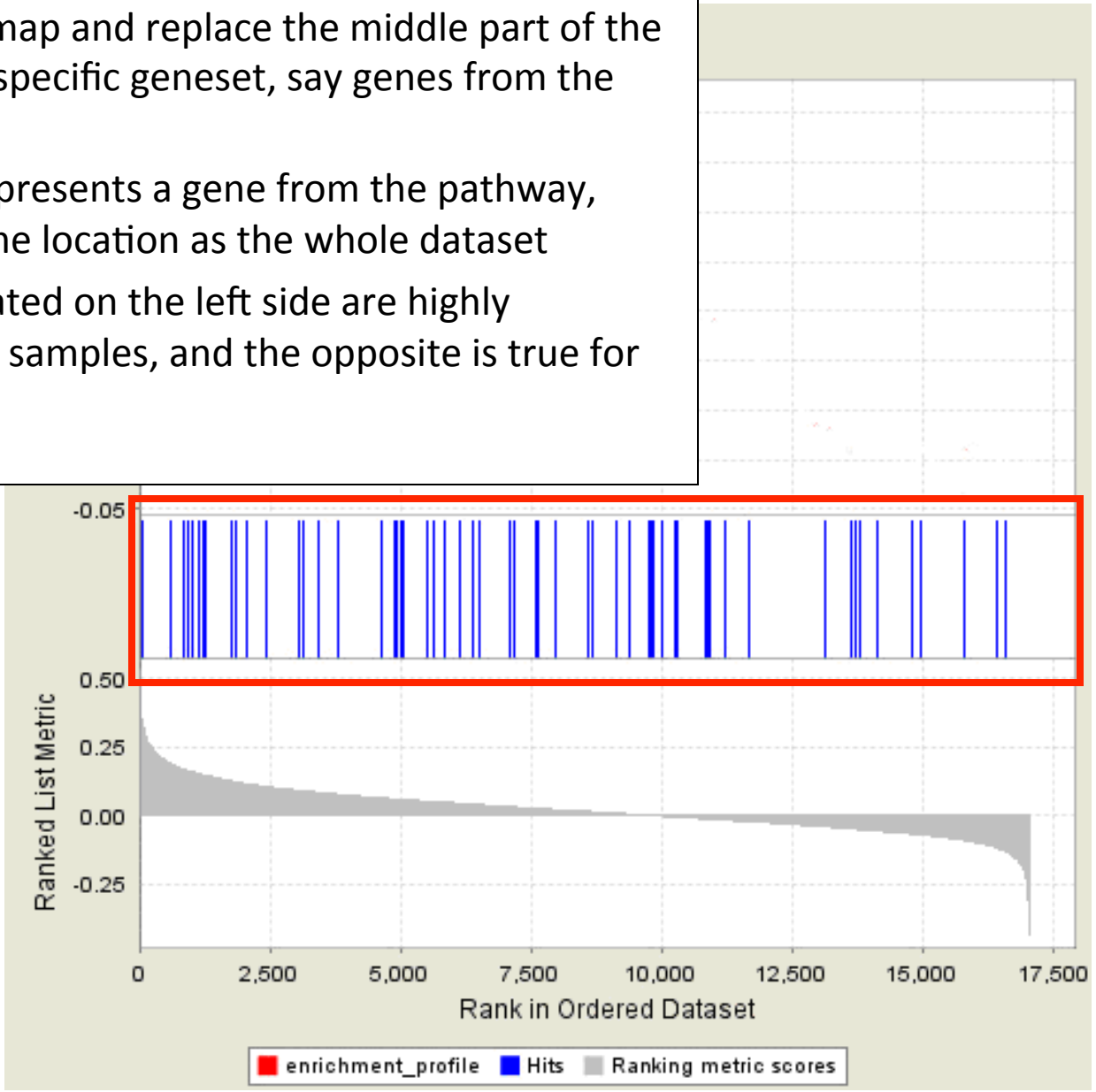
- This is depicted nicely by the graph on the bottom of the figure, where the positive ranks on the left represent the correlation to the Disease phenotype and the negative ranks on the right signify the correlation to the Normal phenotype
- The graph also generates a rank gradient that represents the order of the most up-regulated genes for the Disease sample on the left-most, and the most up-regulated genes for the Normal samples on the right-most

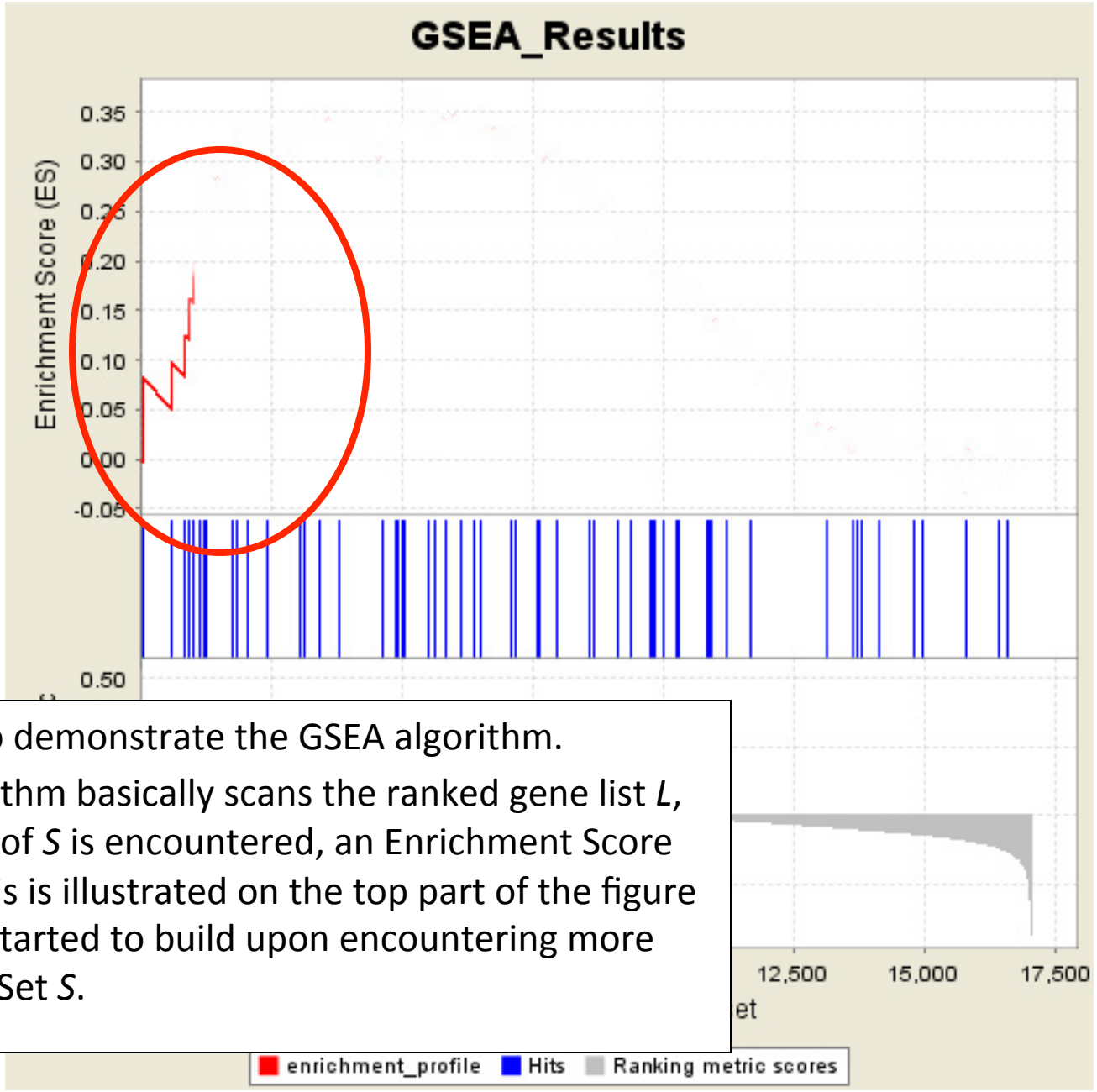
Diseased

Normal

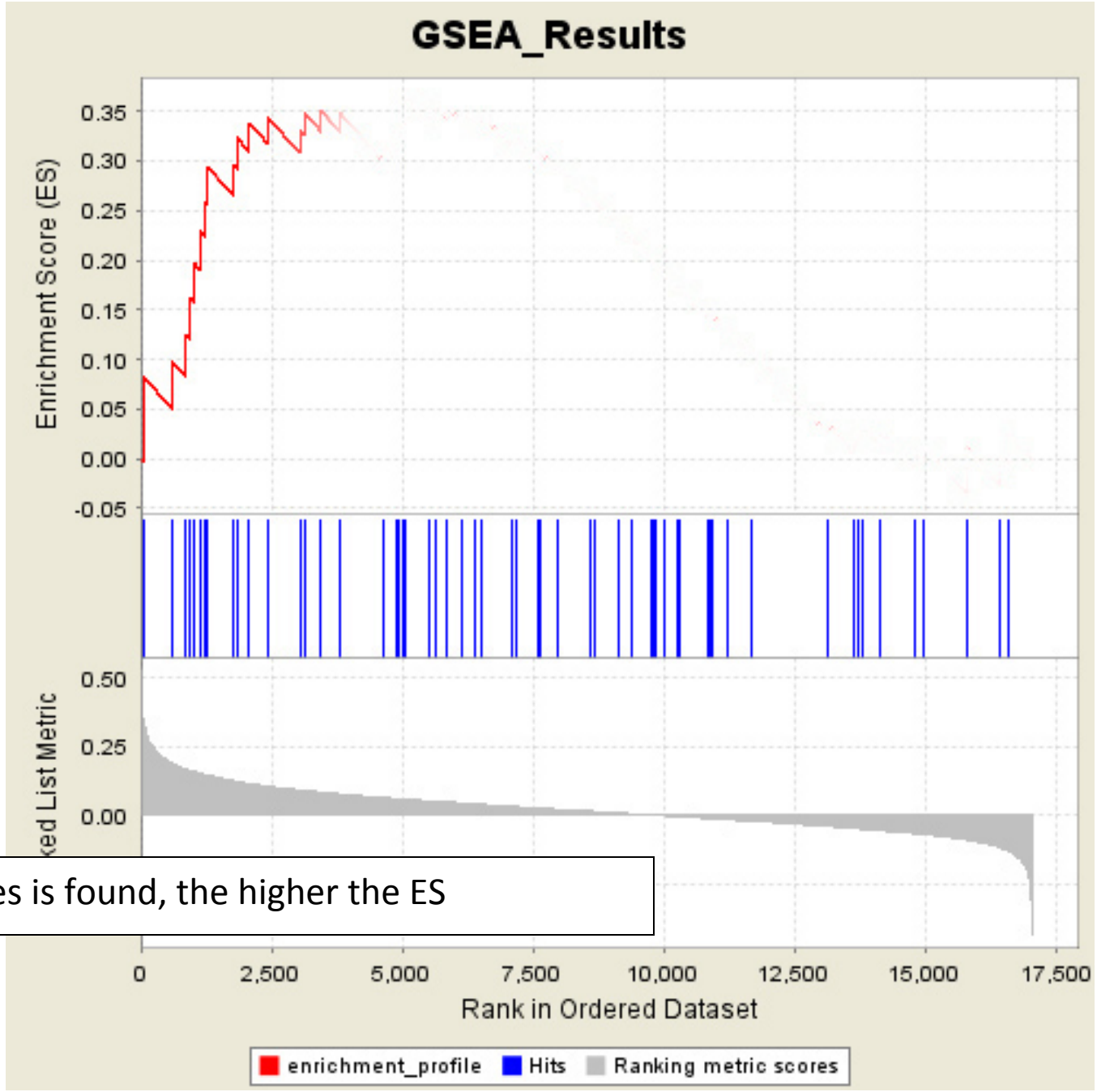


- Now, let's hide the heatmap and replace the middle part of the figure with genes from a specific geneset, say genes from the Glycolysis pathway.
- Each vertical blue bars represents a gene from the pathway, being mapped on the same location as the whole dataset
- Again, genes that are located on the left side are highly expressed on the Disease samples, and the opposite is true for the right-most genes

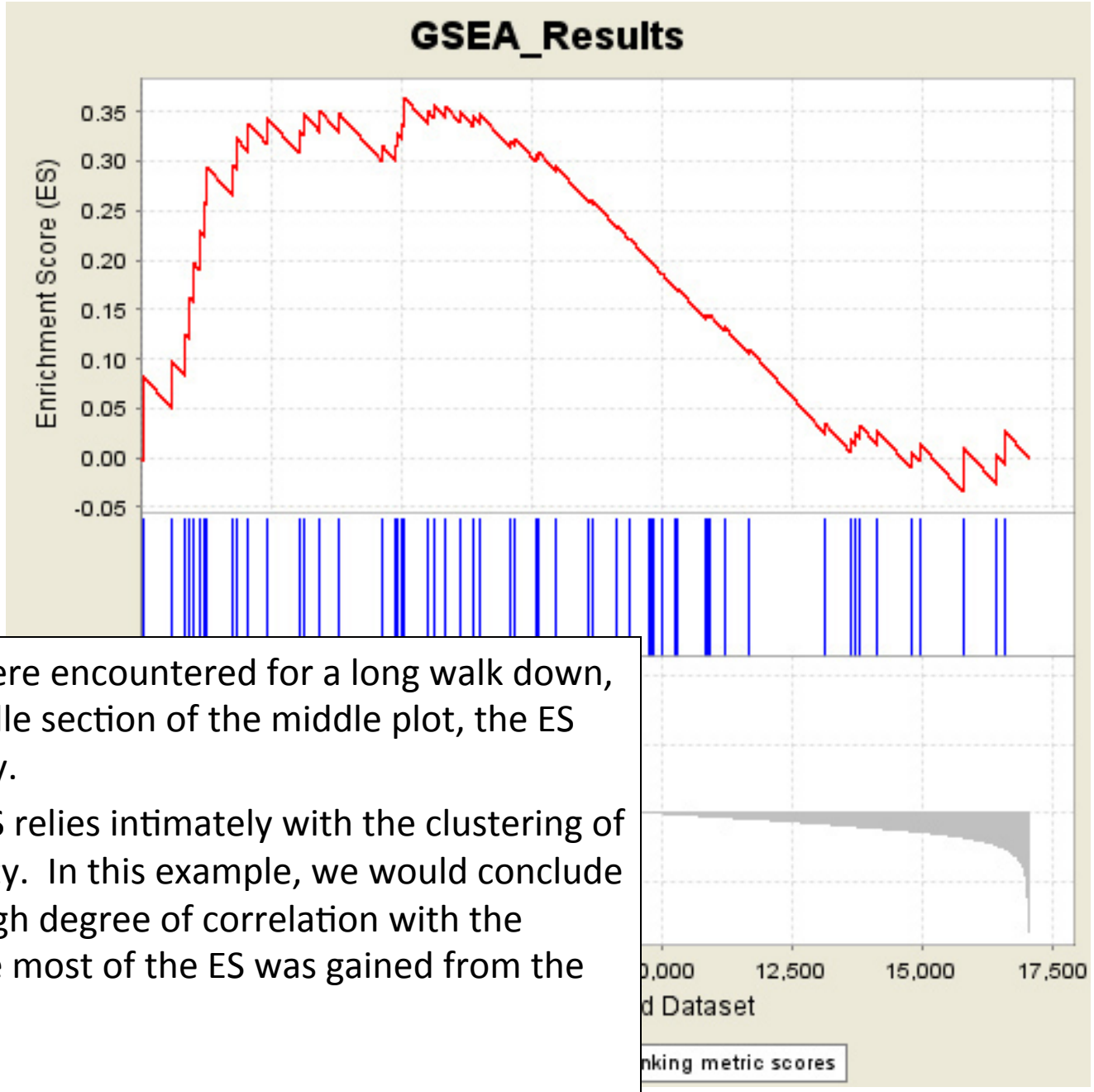




- Now, we are ready to demonstrate the GSEA algorithm.
- The walk down algorithm basically scans the ranked gene list *L*, and when a member of *S* is encountered, an Enrichment Score (*ES*) is registered. This is illustrated on the top part of the figure below; when the ES started to build upon encountering more genes from the GeneSet *S*.



- The more *S* genes is found, the higher the ES

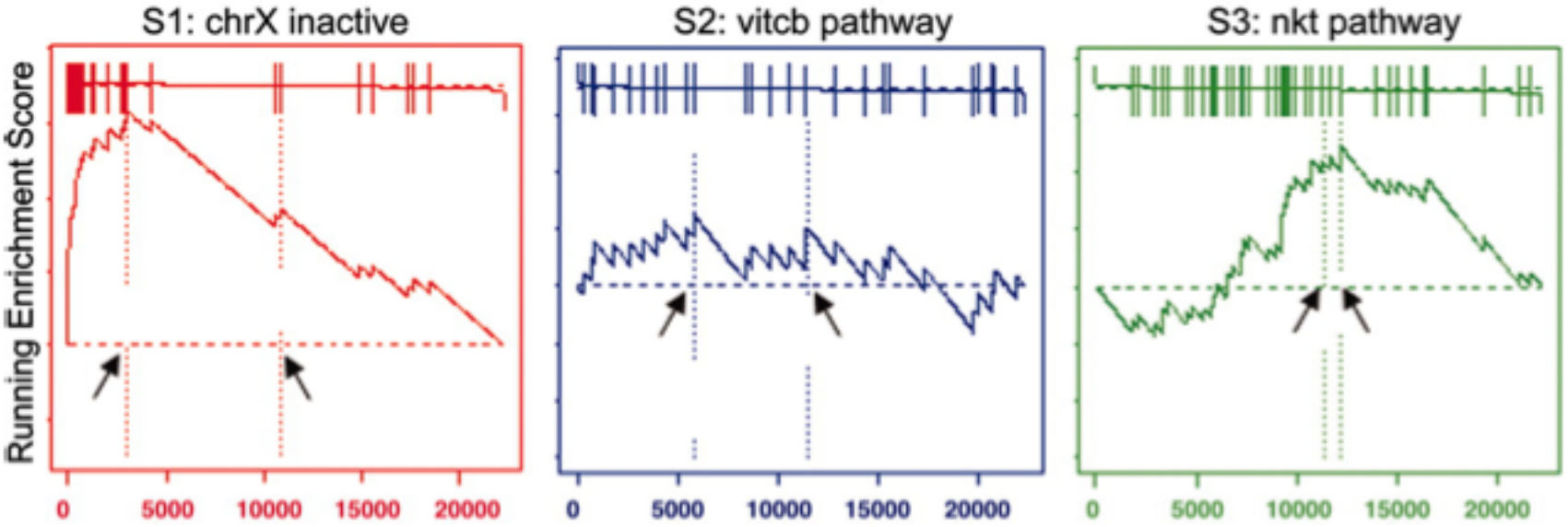


- But, when no *S* genes were encountered for a long walk down, as indicated on the middle section of the middle plot, the ES will decrease accordingly.
- In other words, a high ES relies intimately with the clustering of *S* genes in close proximity. In this example, we would conclude that the *S* genes have high degree of correlation with the Disease phenotype since most of the ES was gained from the left portion of the plot

GSEA Algorithm: Step 1

- Calculate an Enrichment Score:
  - Rank genes by their expression difference
  - Compute cumulative sum over ranked genes:
    - Increase sum when gene in set, decrease it otherwise
    - Magnitude of increment depends on correlation of gene with phenotype
- Record the maximum deviation from zero as the enrichment score

GSEA Algorithm: Step 1



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA Algorithm: Step 2

- Assess significance:
  - Permute phenotype labels 1000 times
  - Compute ES score as above for each permutation
  - Compare ES score for actual data to distribution of ES scores from permuted data

- Permuting the phenotype labels instead of the genes maintains the complex correlation structure of the gene expression data

GSEA Algorithm: Step 3

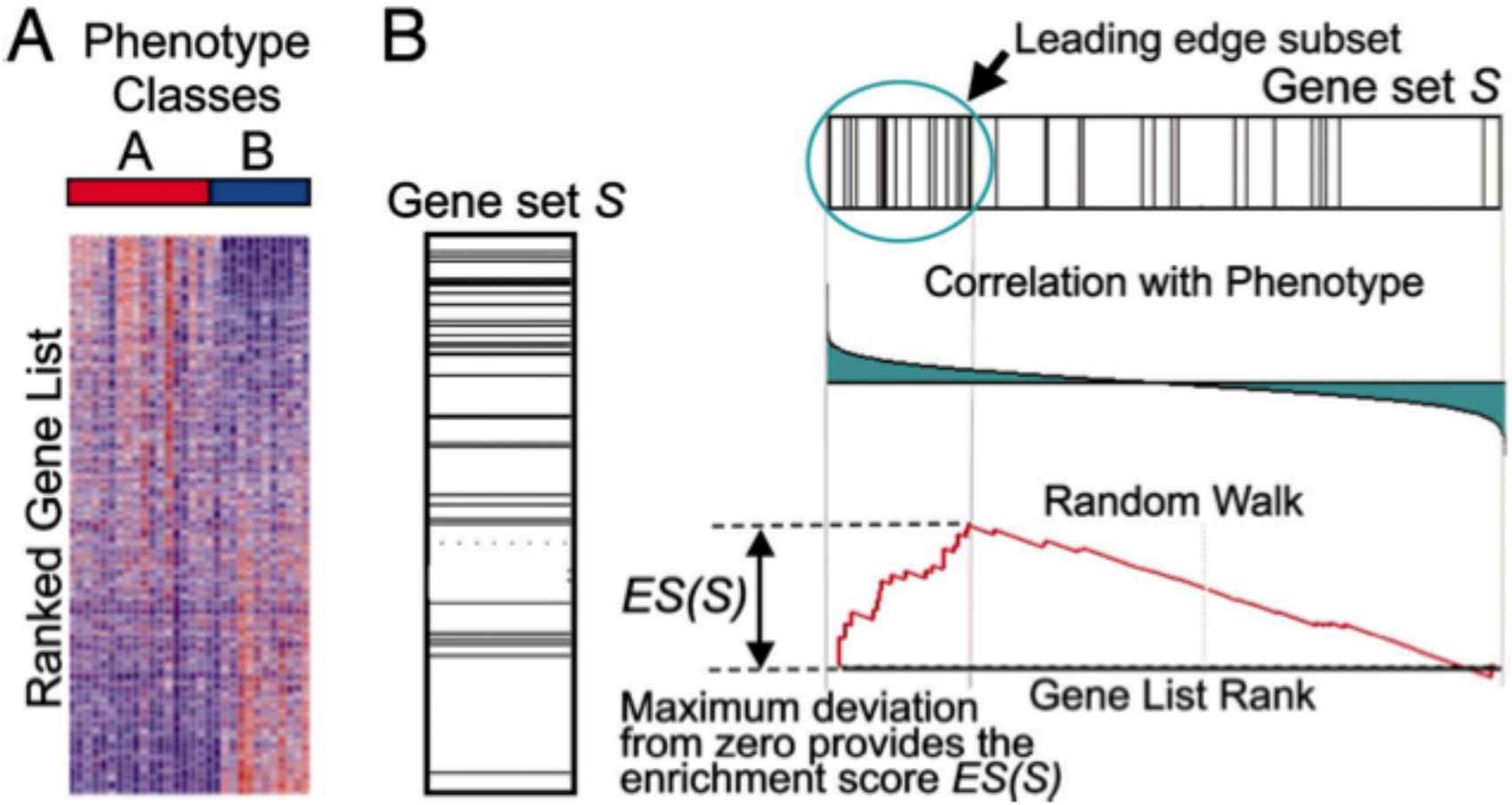
- Adjustment for multiple hypothesis testing:
  - Normalize the ES accounting for size of each gene set, yielding normalized enrichment score (NES)
  - Control proportion of false positives by calculating FDR corresponding to each NES, by comparing tails of the observed and null distributions for the NES

GSEA Algorithm: Step 4

- The original method used equal weights for each gene
  - The revised method weighted genes according to their correlation with phenotype
  - This may cause an asymmetric distribution of ES scores if there is a big difference in the number of genes highly correlated to each phenotype

- Consequently, the above algorithm is performed twice: one for the positively scoring gene sets and once for the negatively scoring gene sets

Overview of GSEA



Subramanian et al., PNAS 102(43), 15545–15550 (2005).

GSEA results for our data set (using pathway gene sets)

Enrichment in phenotype: lean (10 samples)

- 19 / 44 gene sets are upregulated in phenotype **lean**
- 0 gene sets are significant at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 1 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Enrichment in phenotype: obese (9 samples)

- 25 / 44 gene sets are upregulated in phenotype **obese**
- 0 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 12639 native features
- After collapsing features into gene symbols, there are: 6465 genes

Gene set details

- Gene set size filters (min=25, max=500) resulted in filtering out 595 / 639 gene sets
- The remaining 44 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

List of most significant up-regulated gene sets

Table: Gene sets enriched in phenotype lean (10 samples) [plain text format]

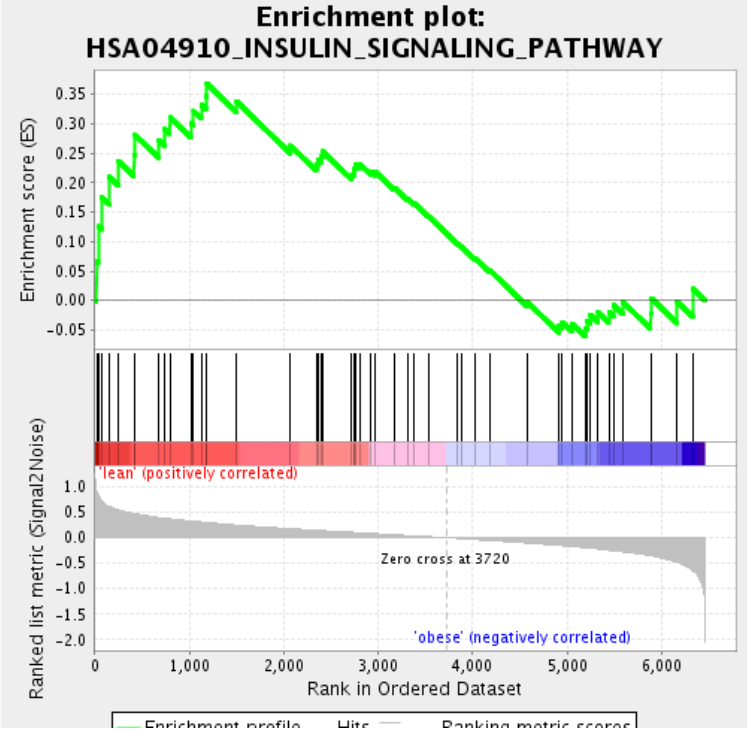
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
1	HSA04910_INSULIN_SIGNALING_PATHWAY	Details ...	51	0.37	1.41	0.036	0.960	0.620	1184
2	CALCINEURIN_NF_AT_SIGNALING	Details ...	32	0.39	1.33	0.074	0.833	0.800	2413
3	HSA04514_CELL_ADHESION_MOLECULES	Details ...	41	0.36	1.26	0.188	0.805	0.880	2038
4	HSA04310_WNT_SIGNALING_PATHWAY	Details ...	52	0.29	1.13	0.278	1.000	0.970	1086
5	HSA04350_TGF_BETA_SIGNALING_PATHWAY	Details ...	29	0.33	1.11	0.302	1.000	0.970	647
6	HSA05215_PROSTATE_CANCER	Details ...	28	0.38	1.11	0.291	0.914	0.970	1360
7	HSA04010_MAPK_SIGNALING_PATHWAY	Details ...	73	0.28	1.03	0.477	1.000	0.990	1482

Table: GSEA Results Summary

Dataset	Pimaunlog2_collapsed_to_symbols.Pima
Phenotype	Pima.cls
Upregulated in class	lean
GeneSet	HSA04910_INSULIN_SIGNALING_PATHWAY
Enrichment Score (ES)	0.3685702
Normalized Enrichment Score (NES)	1.4148982
Nominal p-value	0.035714287
FDR q-value	0.96008533
FWER p-Value	0.62

The Enrichment score is based on the difference of the cumulative distribution of the gene-set minus the expected

This plot is basically the Kolmogorov-Smirnov plot rotated by 45 degrees



Zoom In on Enrichment Plot

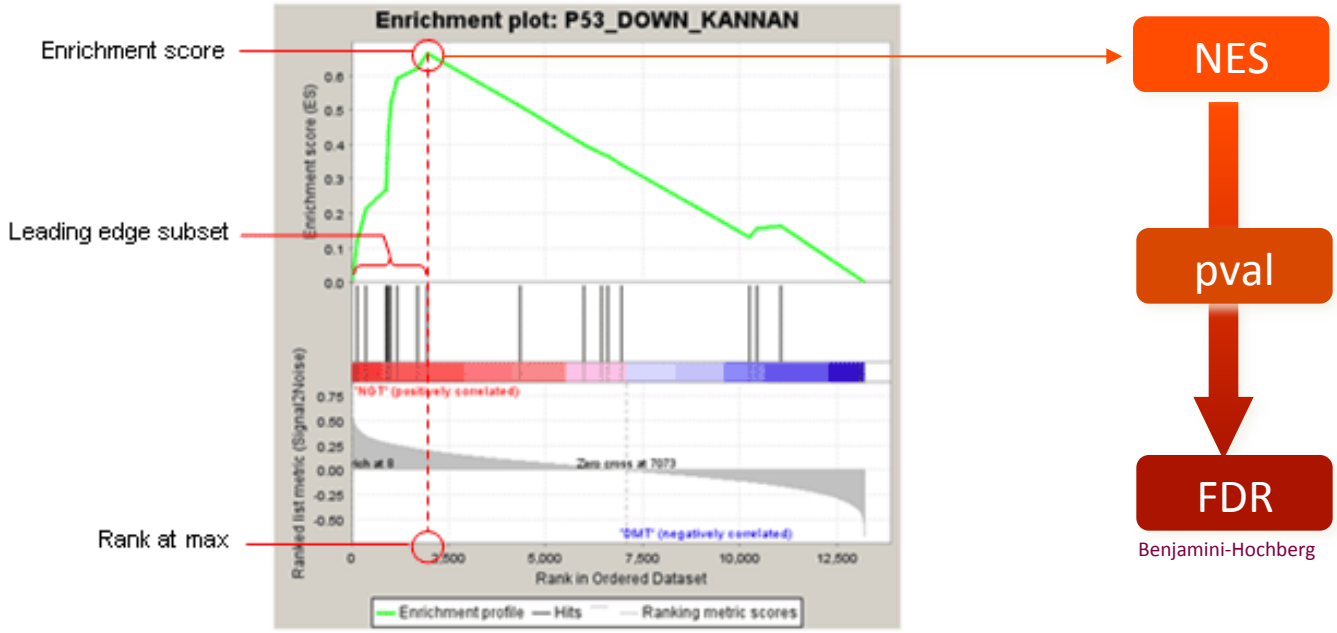
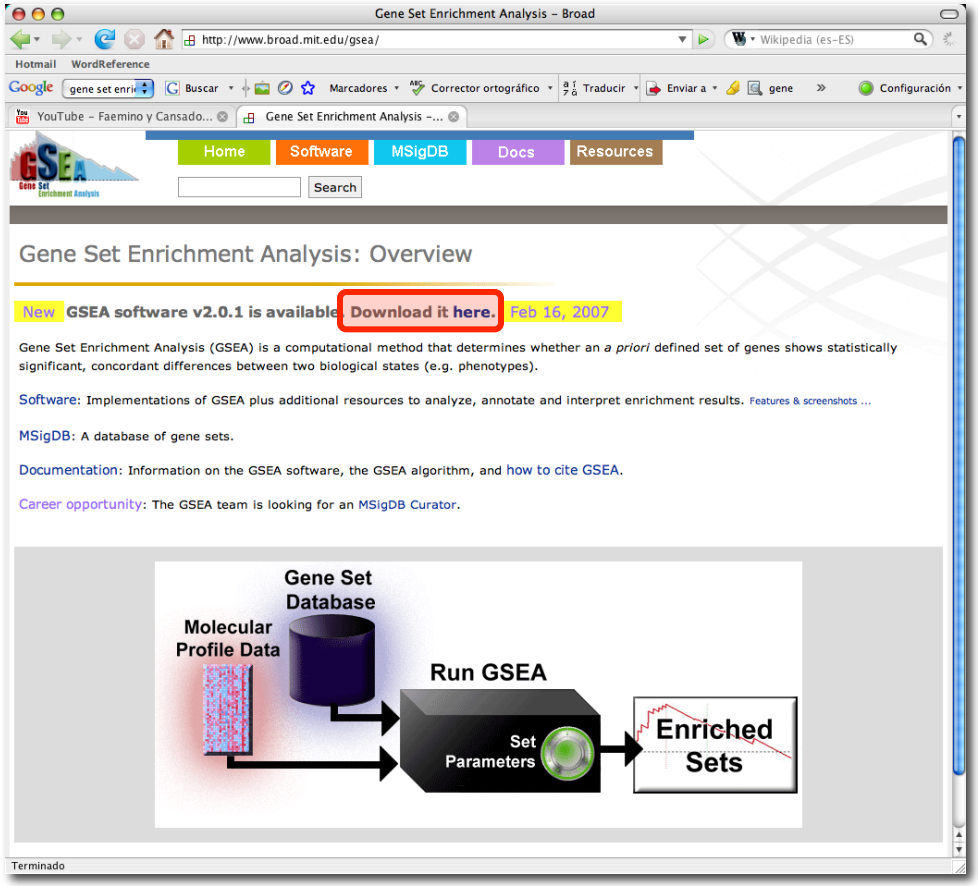


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA Software



http://www.broad.mit.edu/gsea/

Outlook

- Gene Set and Pathway Analysis is a very active field of research: new methods are published all the time!
- One important aspect: taking pathway structure into account
  - All methods we discuss ignored this structure
  - New methods use and "Impact Factor" (IF), which gives more weight to gene that are key regulators in the pathway (Draghici et al (2007))
- Other Aspects:
  - Study the behavior of pathways across experiments in microarray databases like GEO or Array Express
  - Incorporate other data into the analysis (proteomics, metabolomics, sequence data)

Summary

- There are many popular databases/internet resources for pathways and gene sets
- Many important analysis issues
- It is impossible to explain all existing approaches but many of them are some combinations of the methods we discussed
- This is an active field: improvements and further developments are a really active area of research

Questions?