

Pathway and Gene Set Analysis

Part 3

Alison Motsinger-Reif, PhD
Branch Chief, Senior Investigator
Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences

alison.motsinger-reif@niehs.nih.gov

Goals

Some methods in more detail

- TopGO
- Global Ancova
- Pathvisio/Genmapp
- 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*

Newer Reviews

REVIEW ARTICLE

AMERICAN JOURNAL OF
medical genetics PAR
Neuropsychiatric Genetics B

Gene Set Analysis: A Step-By-Step Guide

Michael A. Mooney^{1,2} and Beth Wilmot^{1,2,3*}

¹Department of Medical Informatics & Clinical Epidemiology, Division of Bioinformatics & Computational Biology, Oregon Health & Science University, Portland, Oregon

²OHSU Knight Cancer Institute, Portland, Oregon

³Oregon Clinical and Translational Research Institute, Portland, Oregon

Manuscript Received: 16 March 2015; Manuscript Accepted: 20 May 2015

OXFORD

Briefings in Bioinformatics, 17(3), 2016, 393–407

doi: 10.1093/bib/bbv069

Advance Access Publication Date: 4 September 2015

Paper

Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline

Yasir Rahmatallah, Frank Emmert-Streib and Galina Glazko

Corresponding author: Galina Glazko, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA.
Tel: +1-501-603-1759, Fax: +1-501-526-5964; E-mail: gvglazko@uams.edu

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

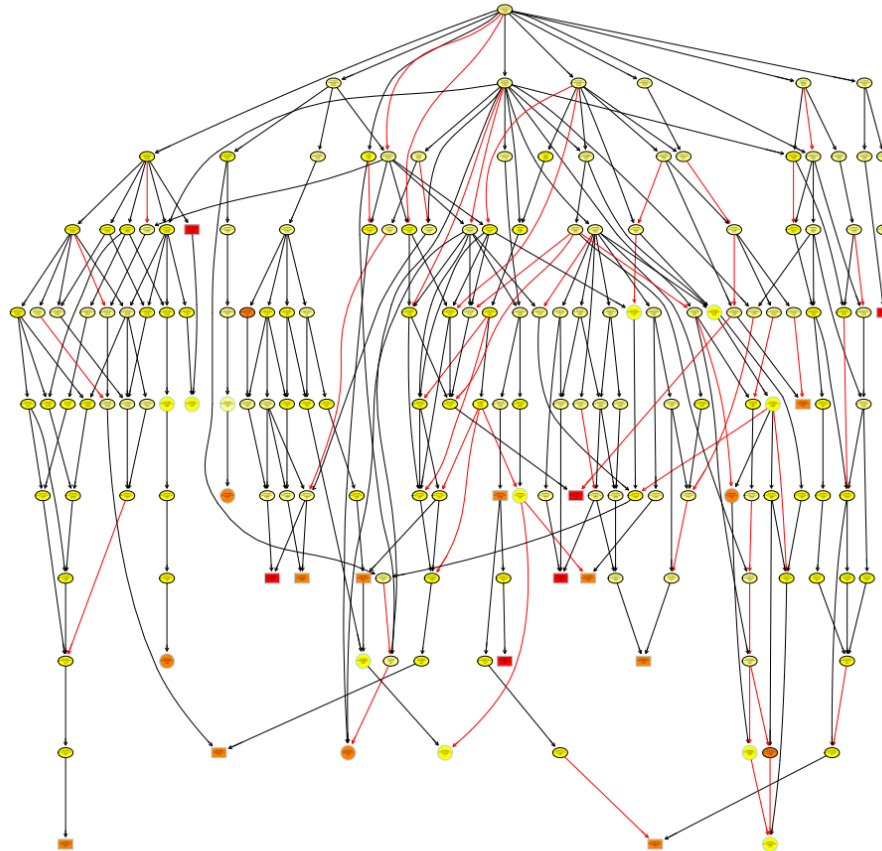
Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006

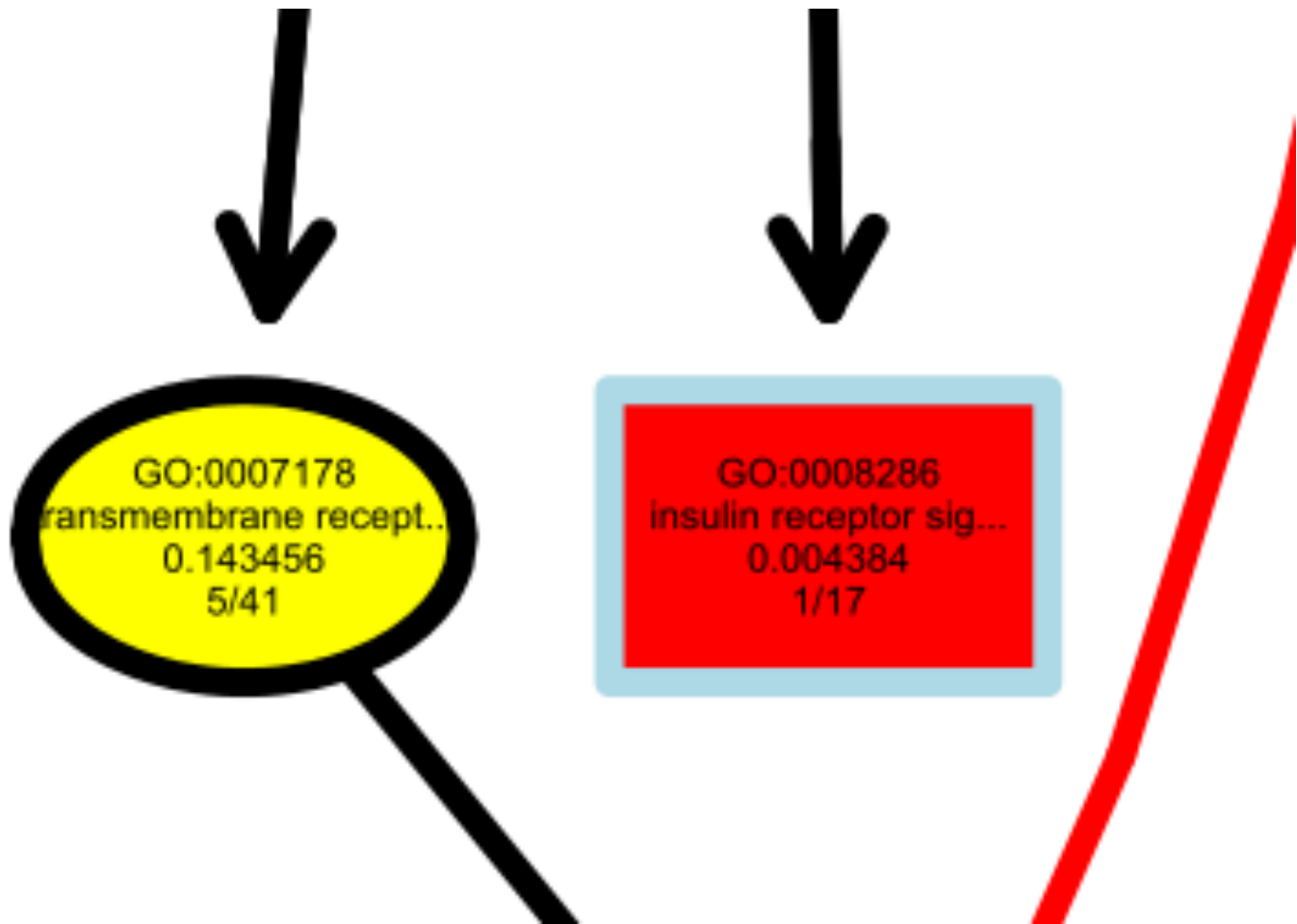
Advance Access publication April 10, 2006

Associate Editor: Martin Bishop

Tree showing the 15 most significant GO terms



Zooming in



Global Ancova

Testing Differential Gene Expression in Functional Groups

Goeman's Global Test versus an ANCOVA Approach

U. Mansmann¹, R. Meister²

¹IBE, Biometry and Bioinformatics, University of Munich, Munich, Germany

²Fachbereich II, University of Applied Sciences, Berlin, Germany

- 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 obese) 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)

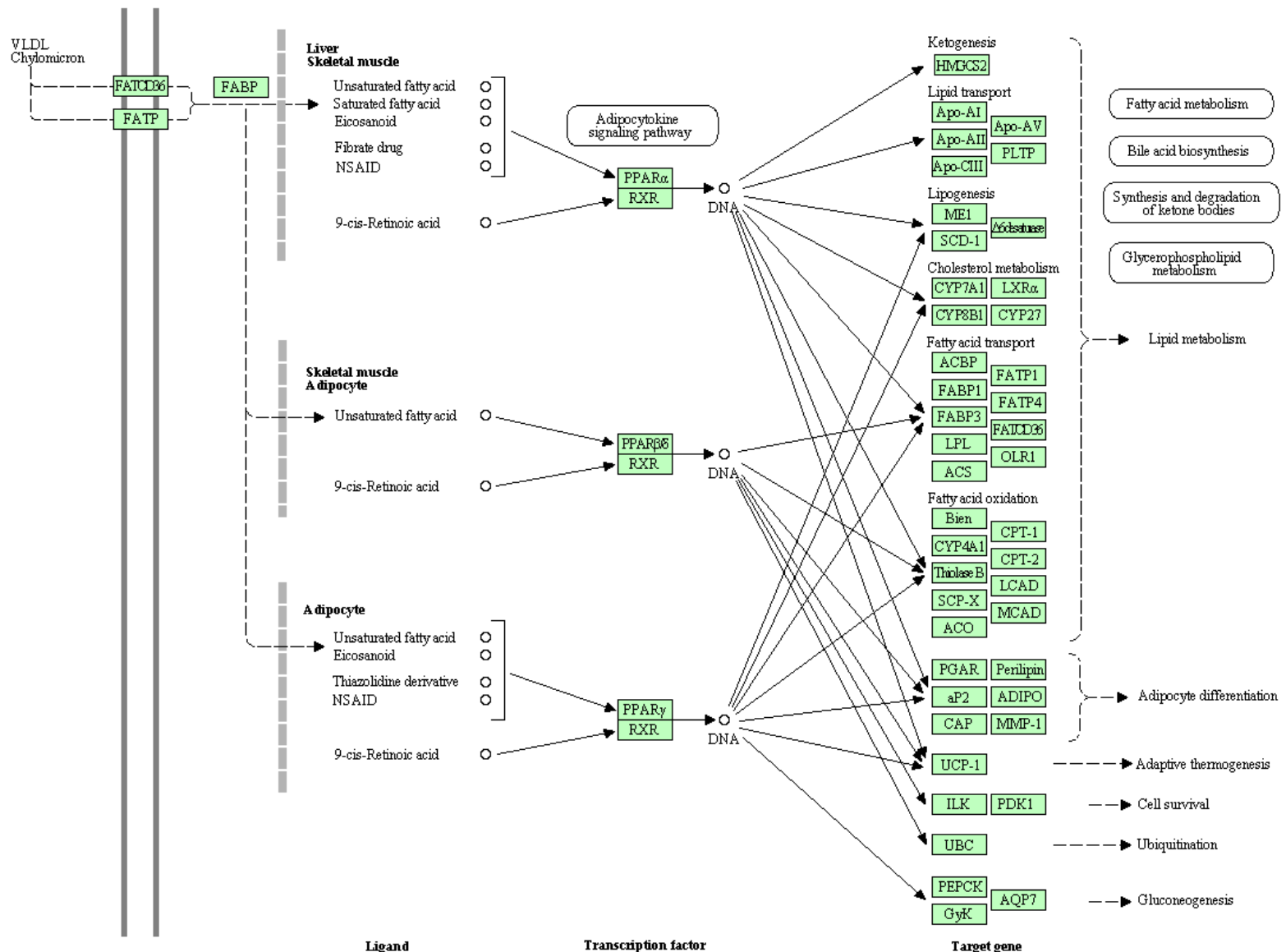
10 most significant KEGG pathways according to Global Ancova

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)

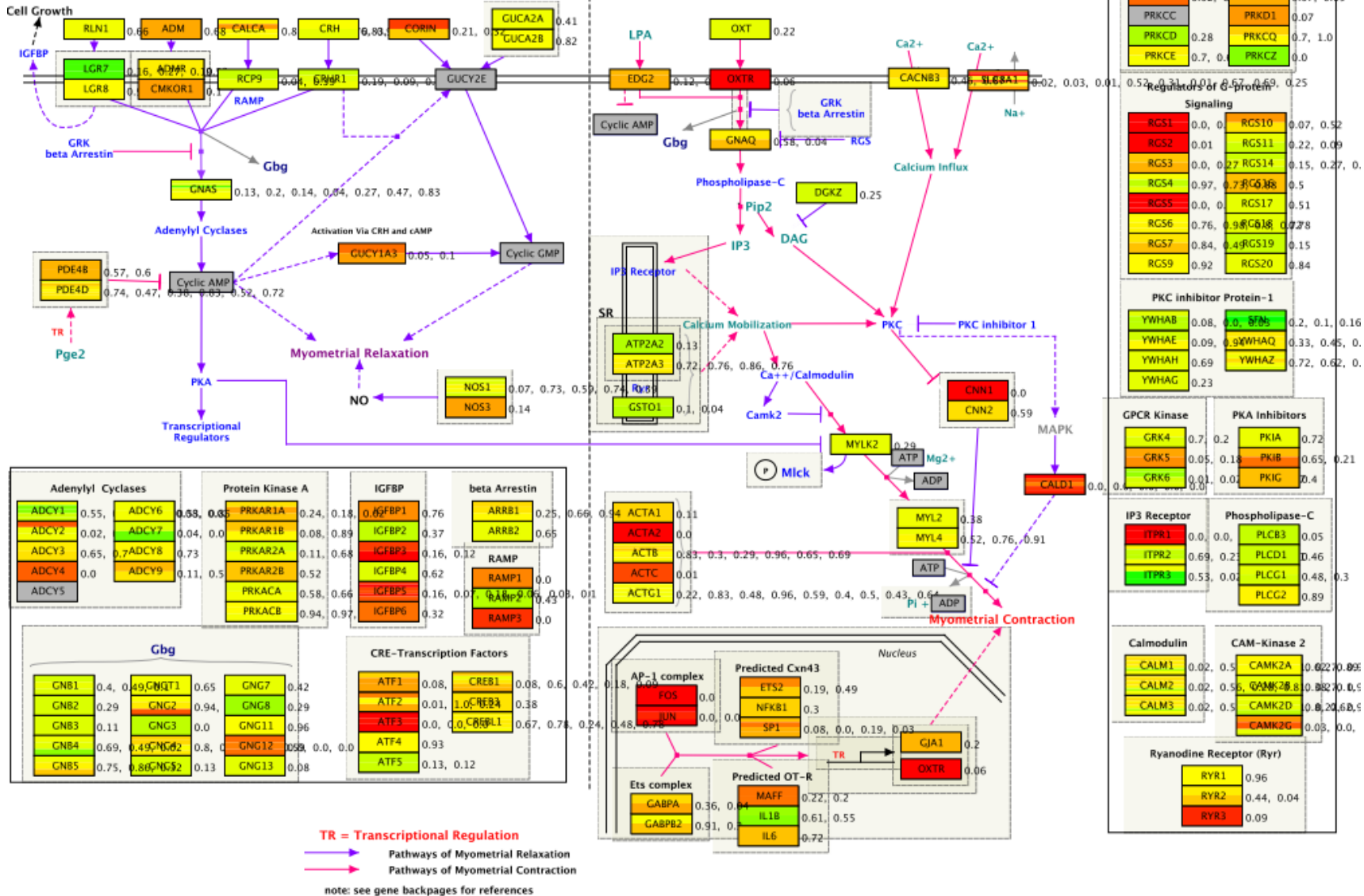
PPAR SIGNALING PATHWAY



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

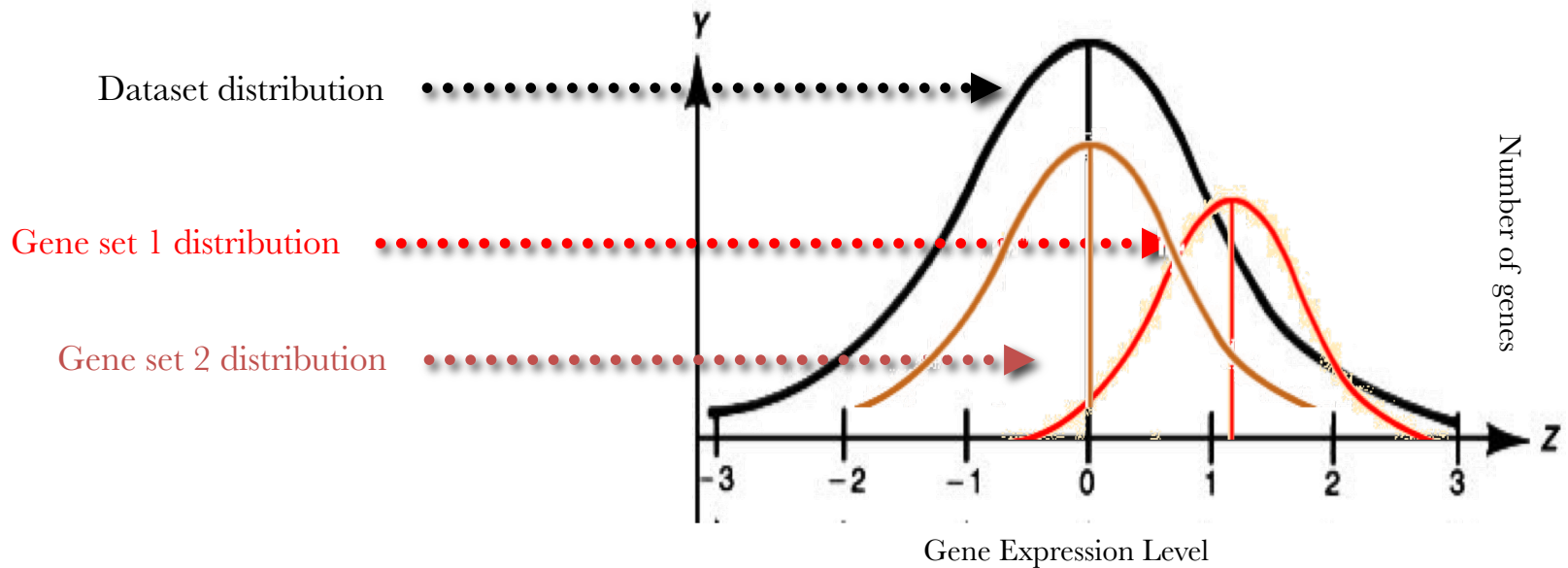


Gene Set Enrichment Analysis (GSEA)

- 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 method 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.

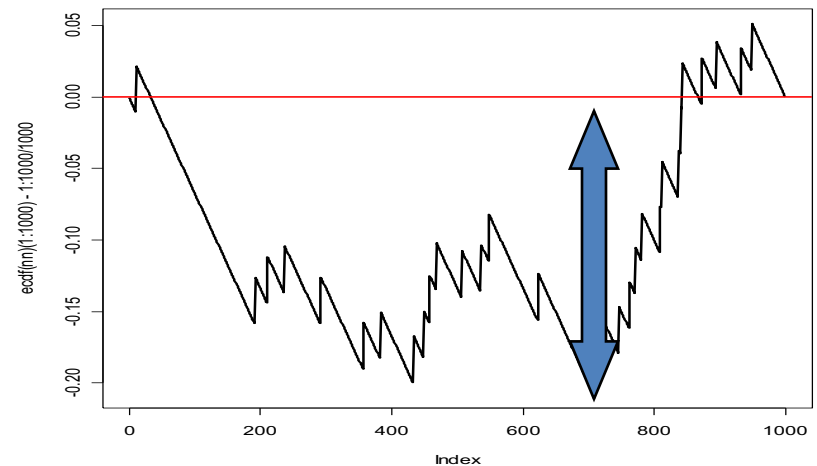
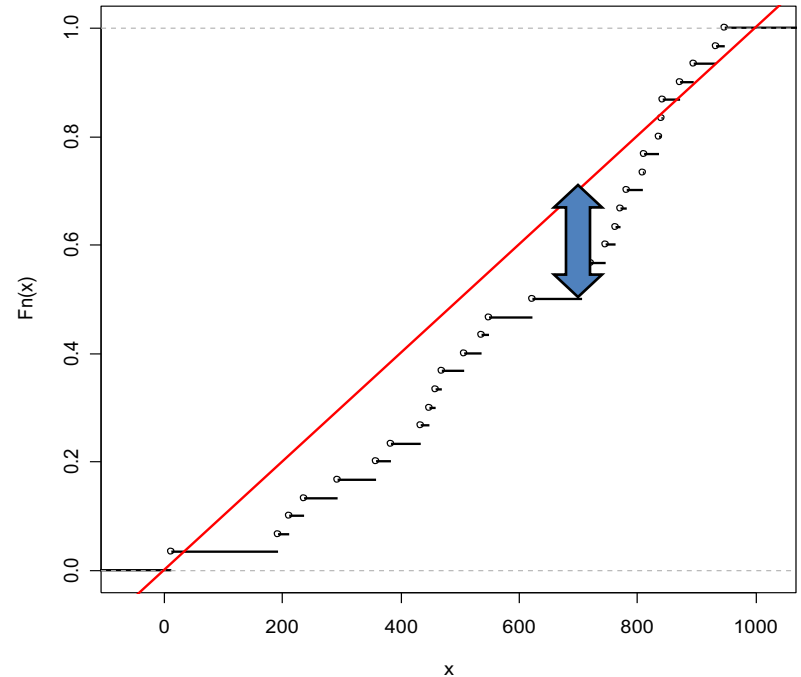
K-S Test

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.

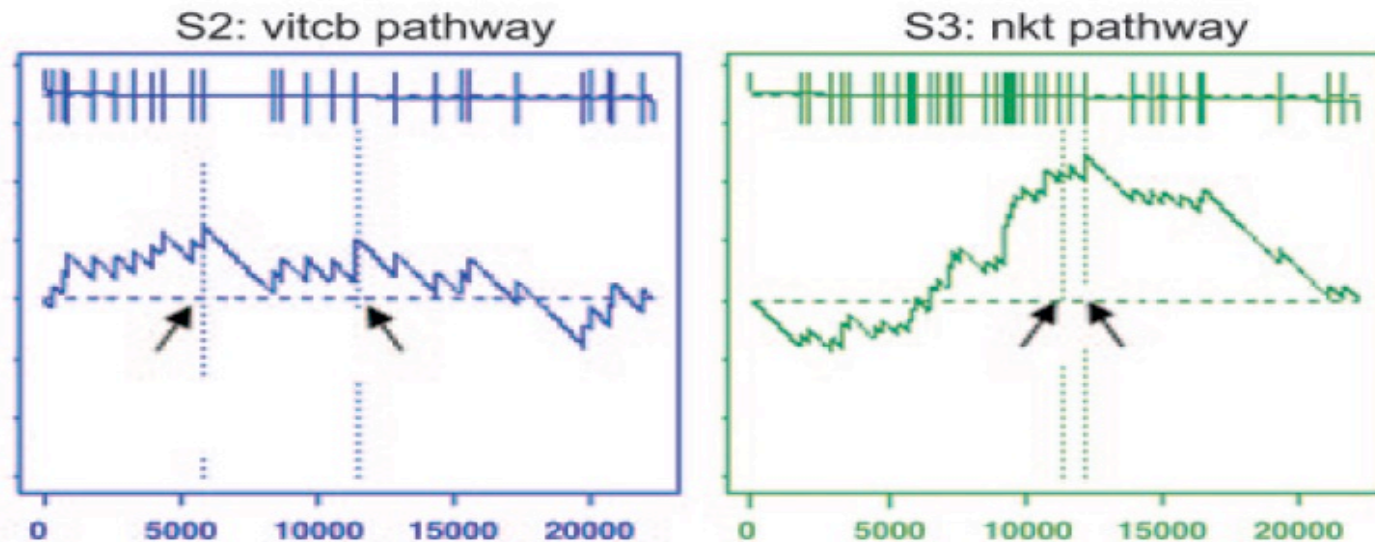


Kolmogorov-Smirnov Test

- 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



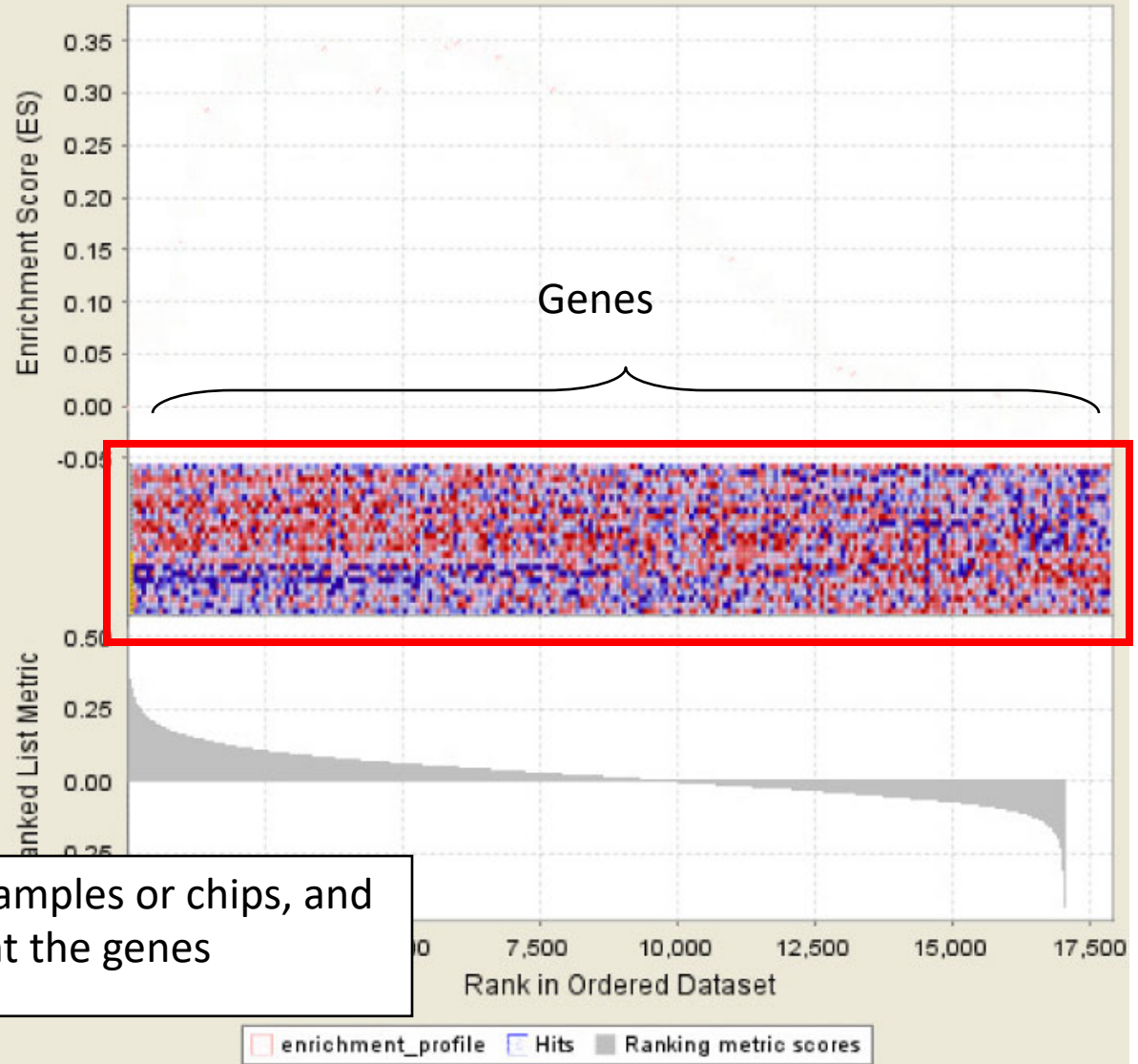
- Sometimes ranks concentrated in middle
 - K-S statistic high, but not meaningful for path change
- Fix: ad-hoc weighting by actual t-scores emphasizes departures at extreme ends
- No theory
- Generate null distribution by permutation

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

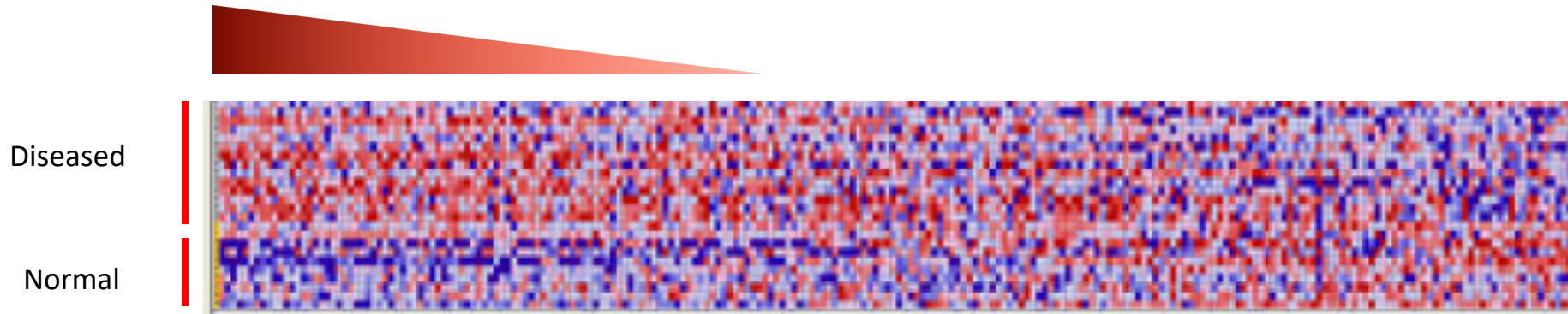
GSEA_Results

Samples



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

Highly expressed in diseased



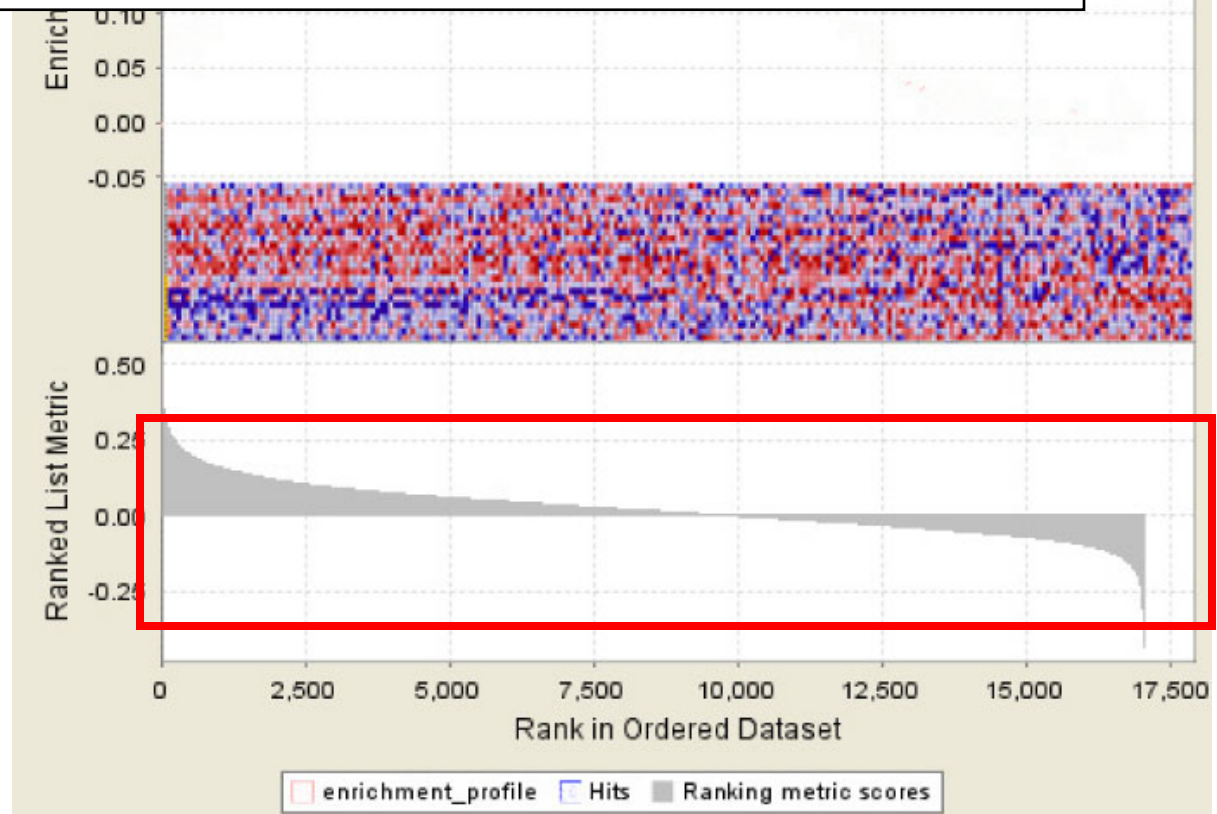
- Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes
- Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes

Lowly expressed in diseased

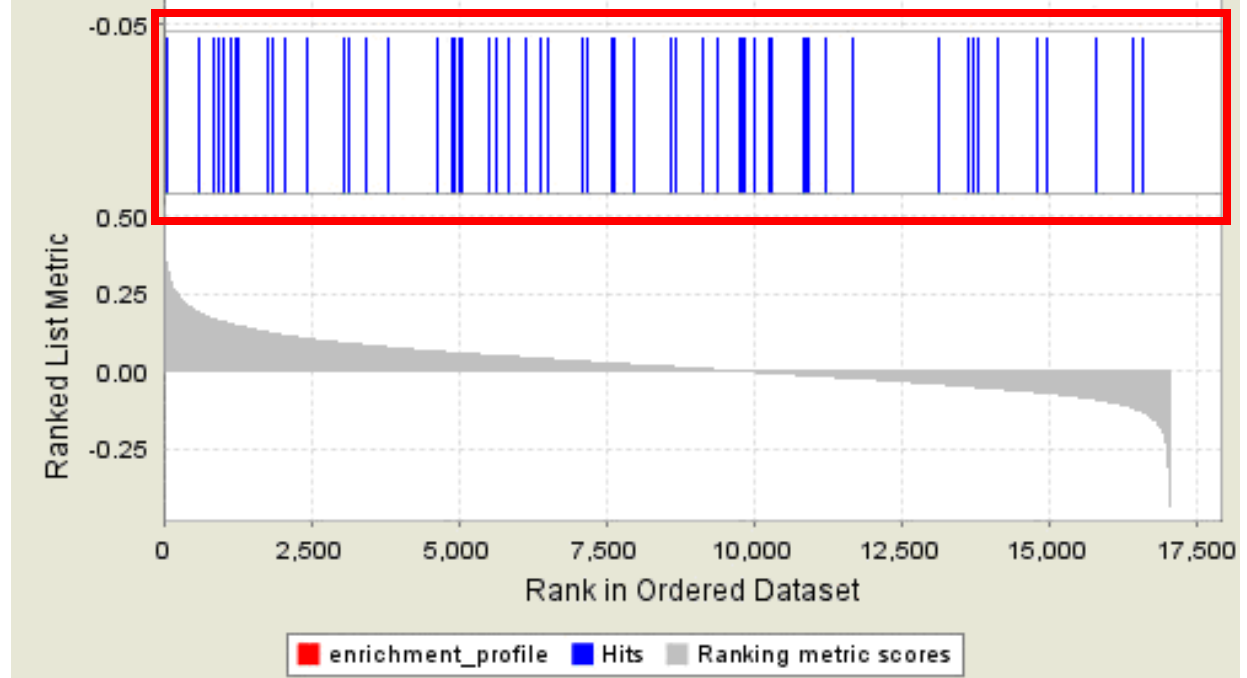
- This is depicted nicely by the graph on the bottom of the figure, where the positive ranks on the left represent the correlation to the Disease phenotype and the negative ranks on the right signify the correlation to the Normal phenotype
- The graph also generates a rank gradient that represents the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

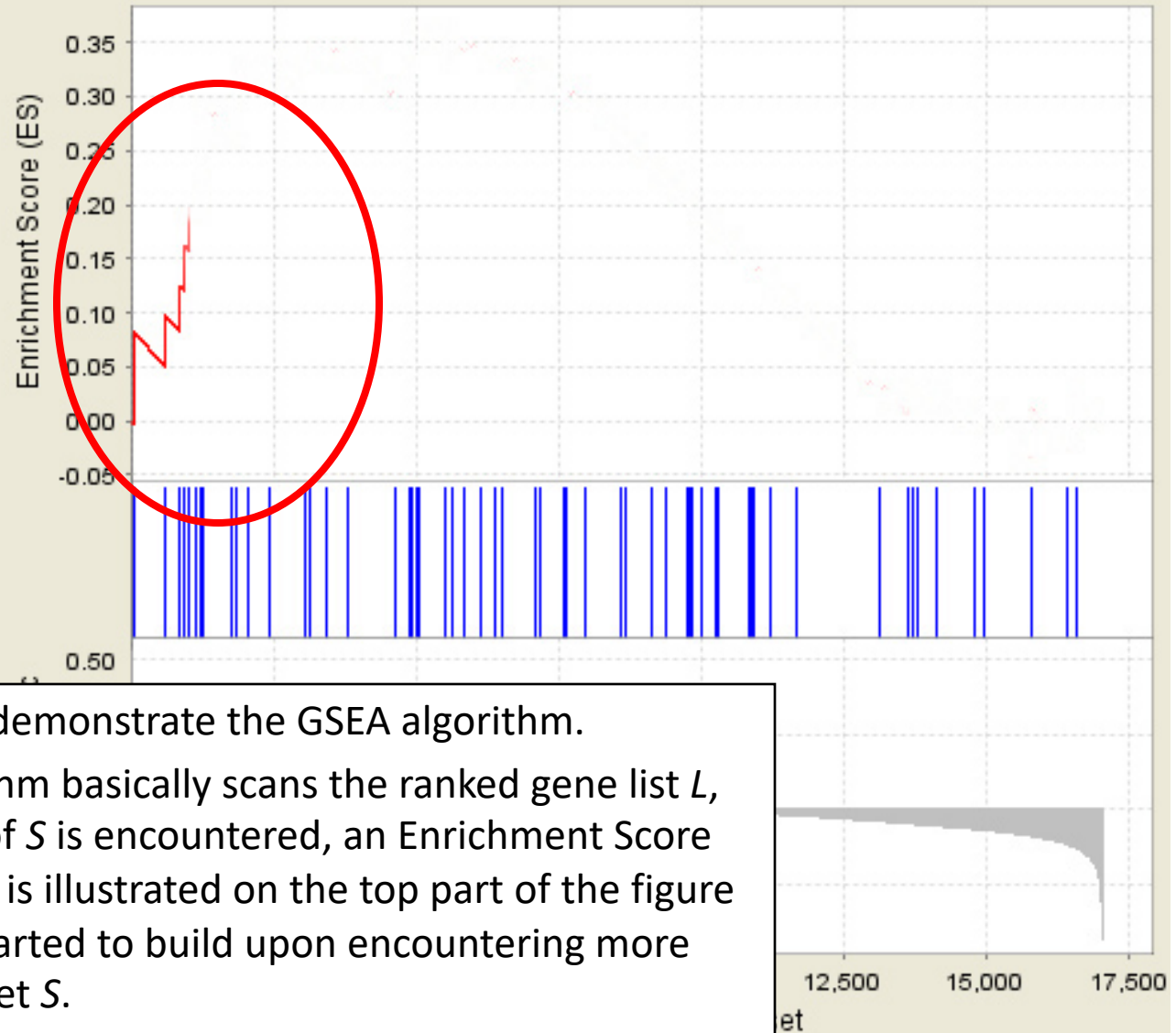
Normal



- 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



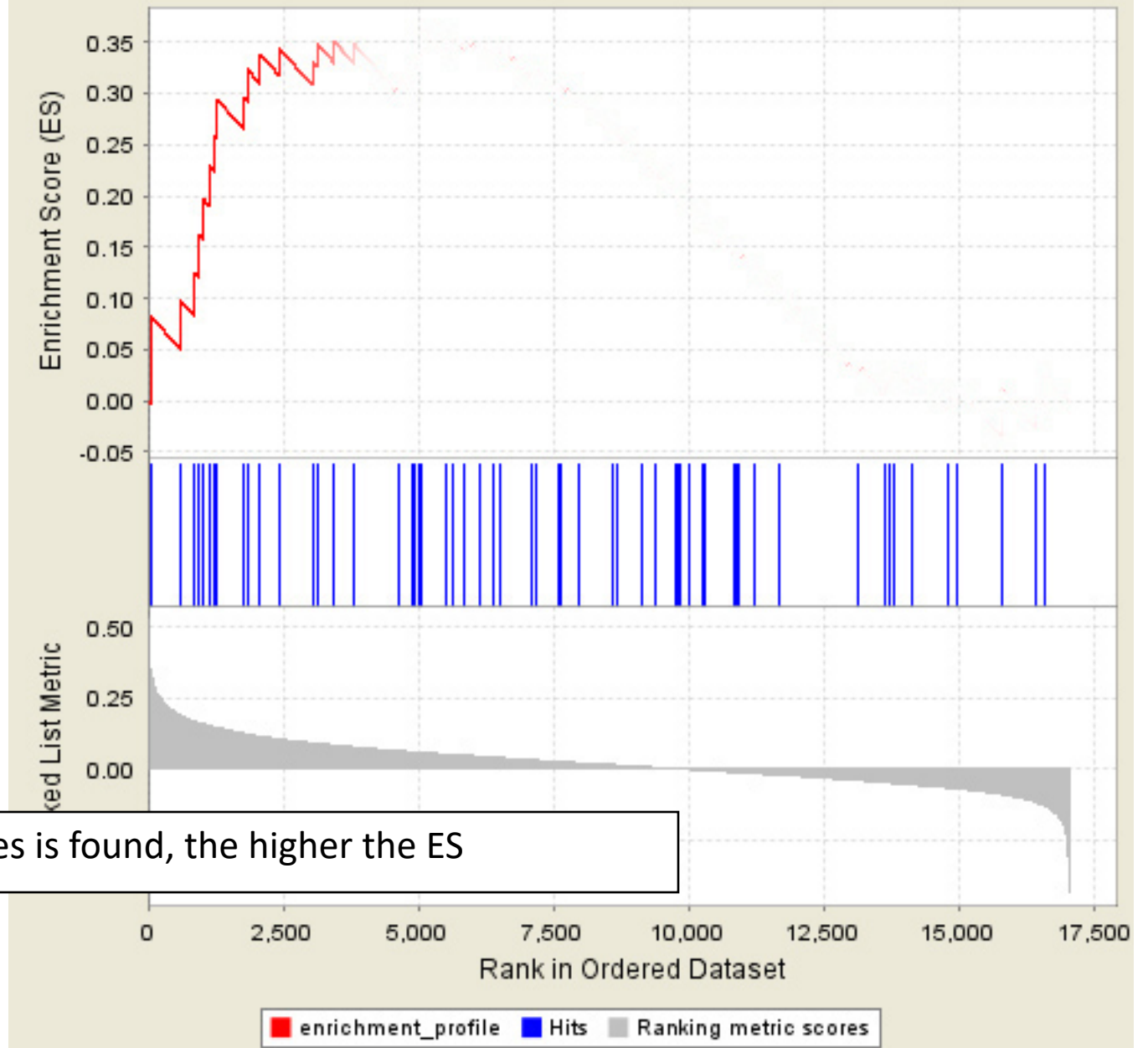
GSEA_Results



- 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 .

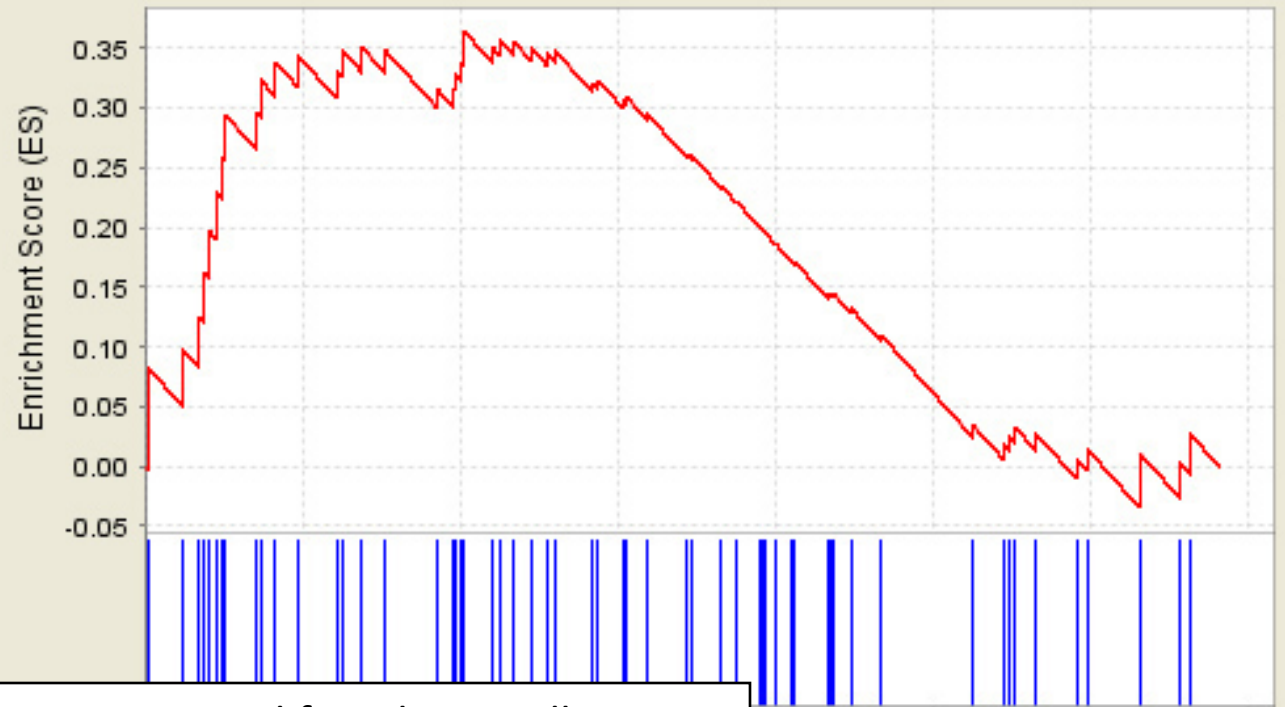
■ enrichment_profile ■ Hits ■ Ranking metric scores

GSEA_Results

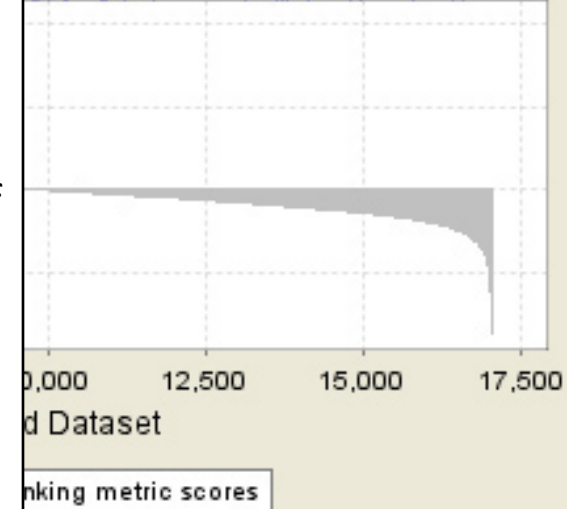


- The more *S* genes is found, the higher the ES

GSEA_Results



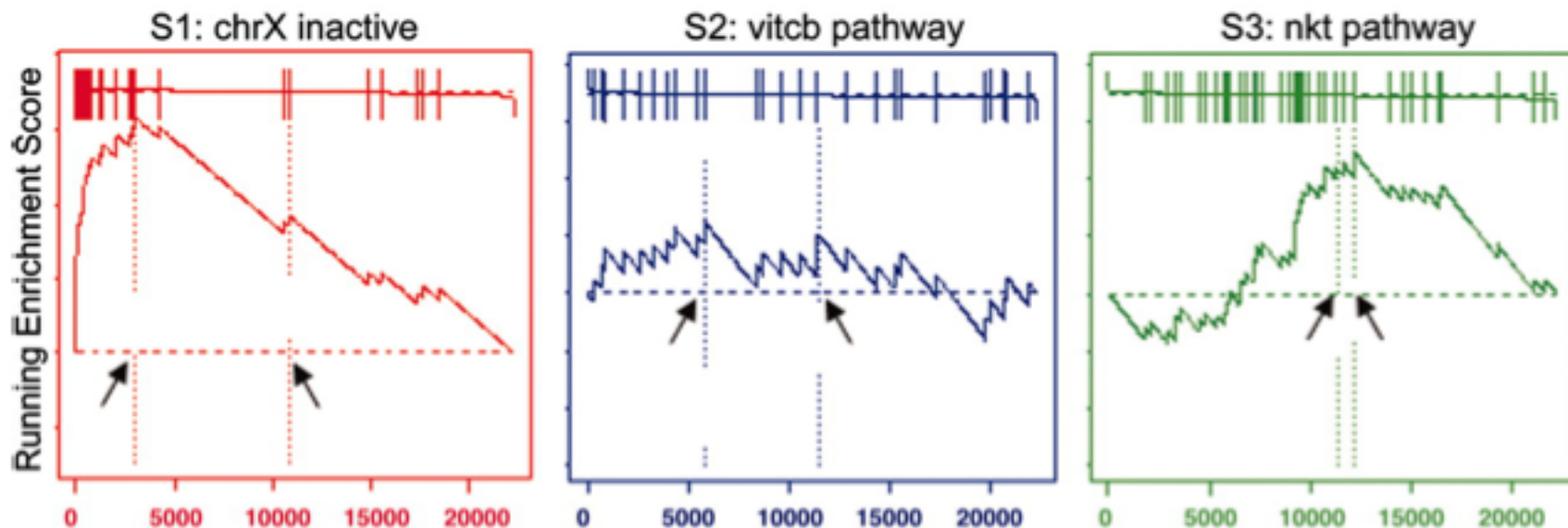
- But, when no *S* genes were encountered for a long walk down, as indicated on the middle section of the middle plot, the ES will decrease accordingly.
- In other words, a high ES relies intimately with the clustering of *S* genes in close proximity. In this example, we would conclude that the *S* genes 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

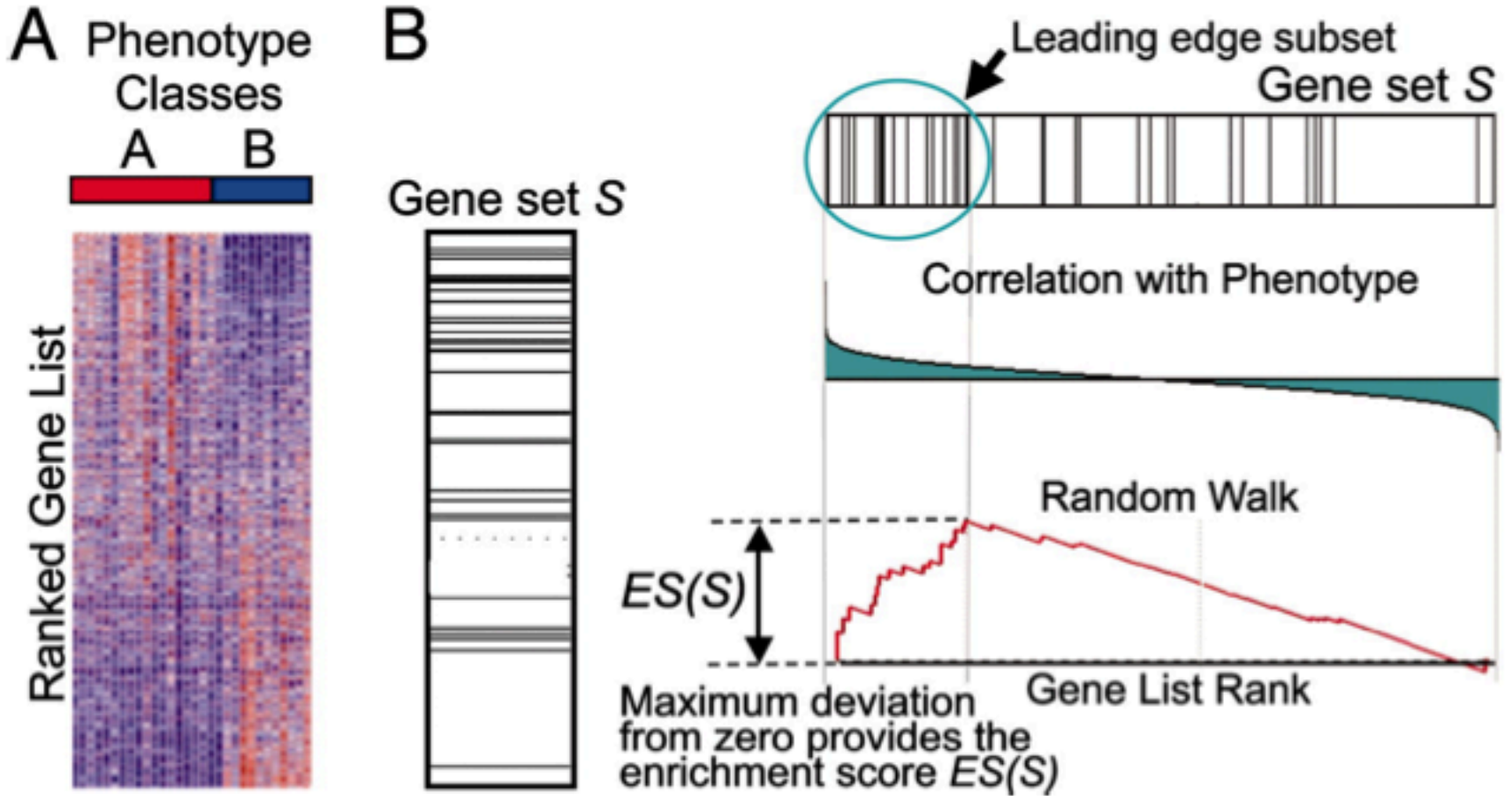
GSEA Algorithm: Step 3

- 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



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

Zoom In on Enrichment Plot

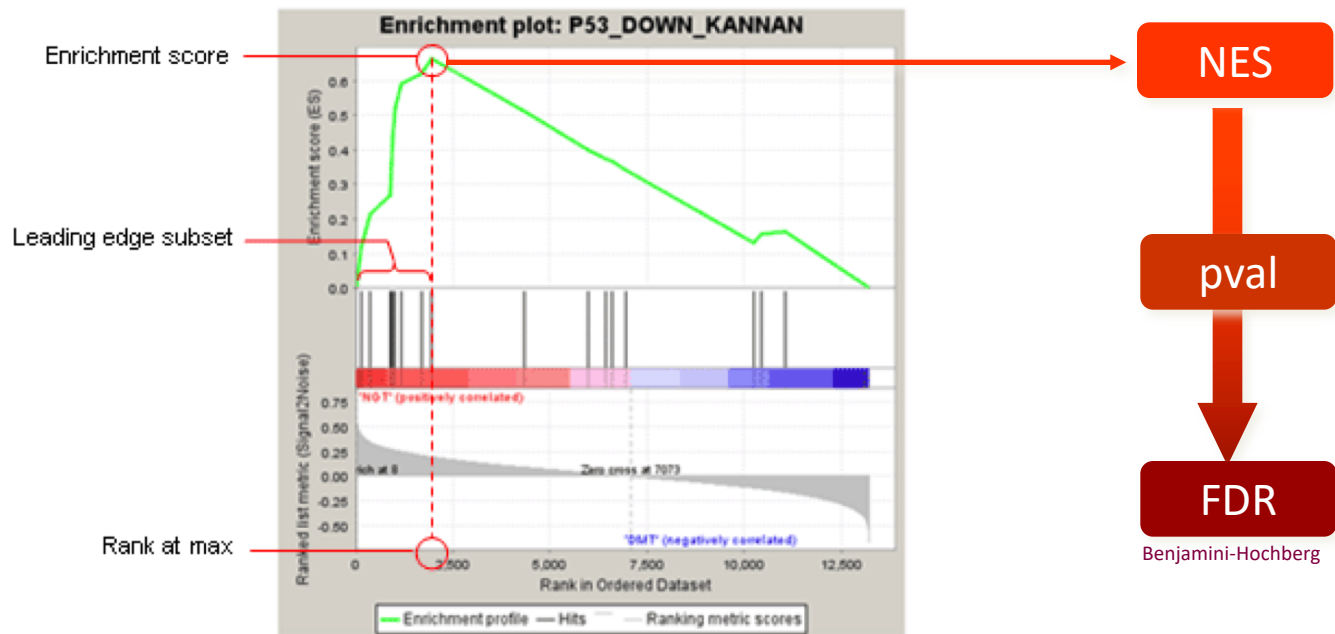


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

GSEA Software

Gene Set Enrichment Analysis - Broad

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

Hotmail WordReference

Google gene set enri: Buscar Marcadores ABC Corrector ortográfico 9/6 Traducir Enviar a gene Configuración

YouTube Faemino y Cansado... Gene Set Enrichment Analysis -...

GSEA Home Software MSigDB Docs Resources

Gene Set Enrichment Analysis: Overview

New GSEA software v2.0.1 is available. Download it here. Jan 16, 2007

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

Software: Implementations of GSEA plus additional resources to analyze, annotate and interpret enrichment results. [Features & screenshots ...](#)

MSigDB: A database of gene sets.

Documentation: Information on the GSEA software, the GSEA algorithm, and [how to cite GSEA](#).

Career opportunity: The GSEA team is looking for an MSigDB Curator.

Diagram: Molecular Profile Data and Gene Set Database feed into Run GSEA (with Set Parameters), resulting in Enriched Sets.

Terminado



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

<https://www.youtube.com/watch?v=KY6SS4vRchY>

Summary

- There are many popular databases/internet resources for pathways and gene sets
 - Discussed methods don't yet account for pathway structure
- 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?