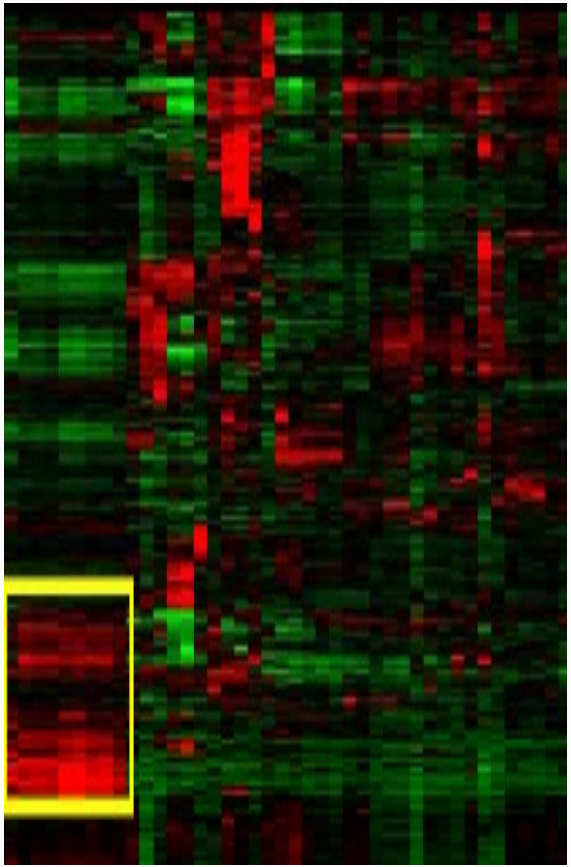
BF528 - Genesets and Enrichment

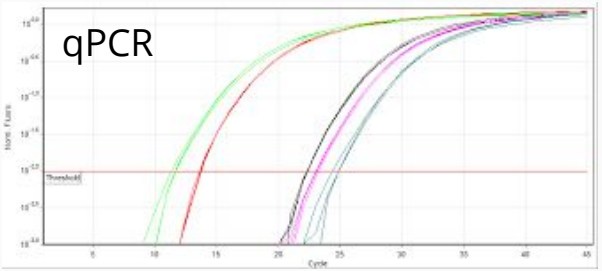
Finding Gene Lists Of Interest

- Gene expression experiments often yield genes implicated in some process
- Most common method: Get a list of differentially expressed genes
  - P-value and/or fold change?
  - Threshold?
- Alternatives:
  - Define a cluster
  - Sort data and/or apply a model to rank genes
- Recommendations:
  - Try lists of varying length
  - Try to maximize signal / noise



You found interesting genes: now what?

- Select some genes for validation
- Do some follow-up experiments
- Publish a huge table with results
- Try to learn about genes from published literature



Gene type	Gene symbol[a]	Entrez gene ID	Accession number	Gene name	Function of encoded protein[b]	TaqMan assay	Exon	Assay location[f]	Amplicon length (bp)
Candidate reference genes	*CHCHD1*	118487	NM_203298.2	Coiled-coil-helix-coiled coil-helix domain containing 1	Component of the mitochondrial ribosome small subunit (28S)	Hs00415053_g1[c]	1–2	153	98
	GNB2L1	10399	NM_006098.4	Guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	Possibly involved in protein kinase C (PKC) signaling	Hs00914568_g1[c]	4–5	625	75
	IPO8	10526	NM_006390.3	Importin 8	Involved in nuclear import of proteins	Hs00183533_m1[c]	20–21	2615	71
	LASP1	3927	NM_006148.3	LIM and SH3 protein 1	Actin-binding protein	Hs00196221_m1[c]	6–7	946	82
	RPL27A	6157	NM_000990.4	Ribosomal protein L27a	Component of the 60S subunit of the ribosomes	Hs00741143_s1[d]	5–5	4471	94
	RPS12	6206	NM_001016.3	Ribosomal protein S12	Component of the 40S subunit of the ribosomes	Hs00831630_g1[e]	6–6	437	109

Interpreting Gene Lists

- Single gene analysis method instrumental in our understanding of cell-biological process
- Many genes work together in the cell in, e.g. pathways
- Biologically, gene expression of entire cellular pathways is perturbed as a unit
- Individual genes cannot reveal high level patterns
- Groups of genes known to relate are organized into *genesets*

Types of Gene Sets

- Genes are organized in different ways:
  - Biological process (e.g. cell cycle, inflammation)
  - Molecular function (kinase, zinc ion binding)
  - Cellular component (nucleus, cell membrane)
  - Disease process (transcription in cancer)
  - Genome loci (chromosome 1 p32.1)

Does our gene list overlap with a given gene set?
Which gene sets are implicated by our gene list?

Popular Annotation Sources

- KEGG Pathways - Kyoto Encyclopedia of Genes and Genomes
- REACTOME - curated, peer-reviewed pathway DB
- **Gene Ontology**
- Genes sharing a motif or regulated by the same protein/miRNA
- Genes found on the same chromosome
- **Broad's Molecular Signatures Database (MSigDB)**
- any grouping that is biologically sensible (e.g. your own)

Molecular Signatures Database - MSigDB

The Molecular Signatures Database (**MSigDB**) gene sets are divided into 8 major collections: http://software.broadinstitute.org/gsea/msigdb

C1: positional gene sets

C2: curated gene sets

C3: motif gene sets

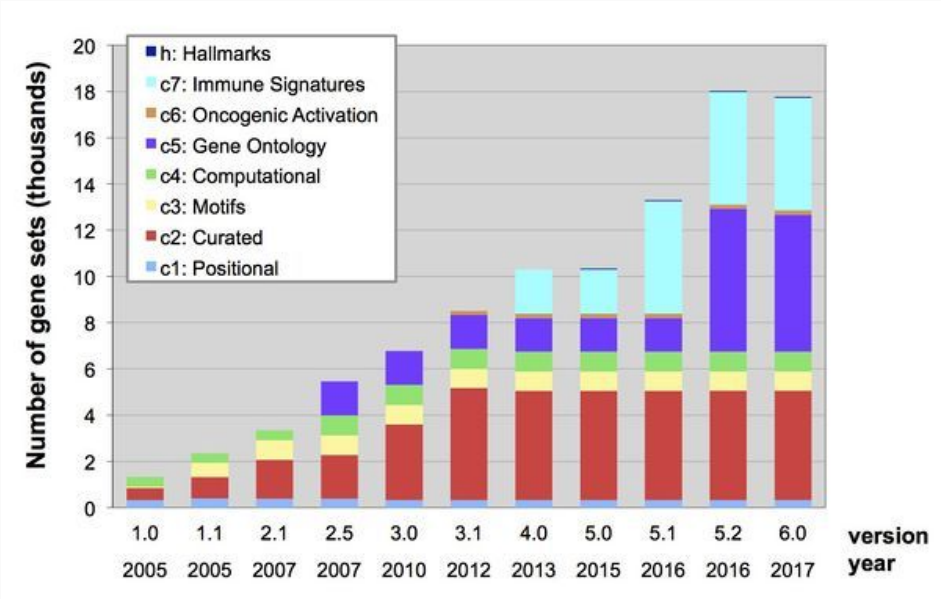
C4: computational gene sets

C5: GO gene sets

C6: Oncogenic signatures

C7: immunogenic signatures

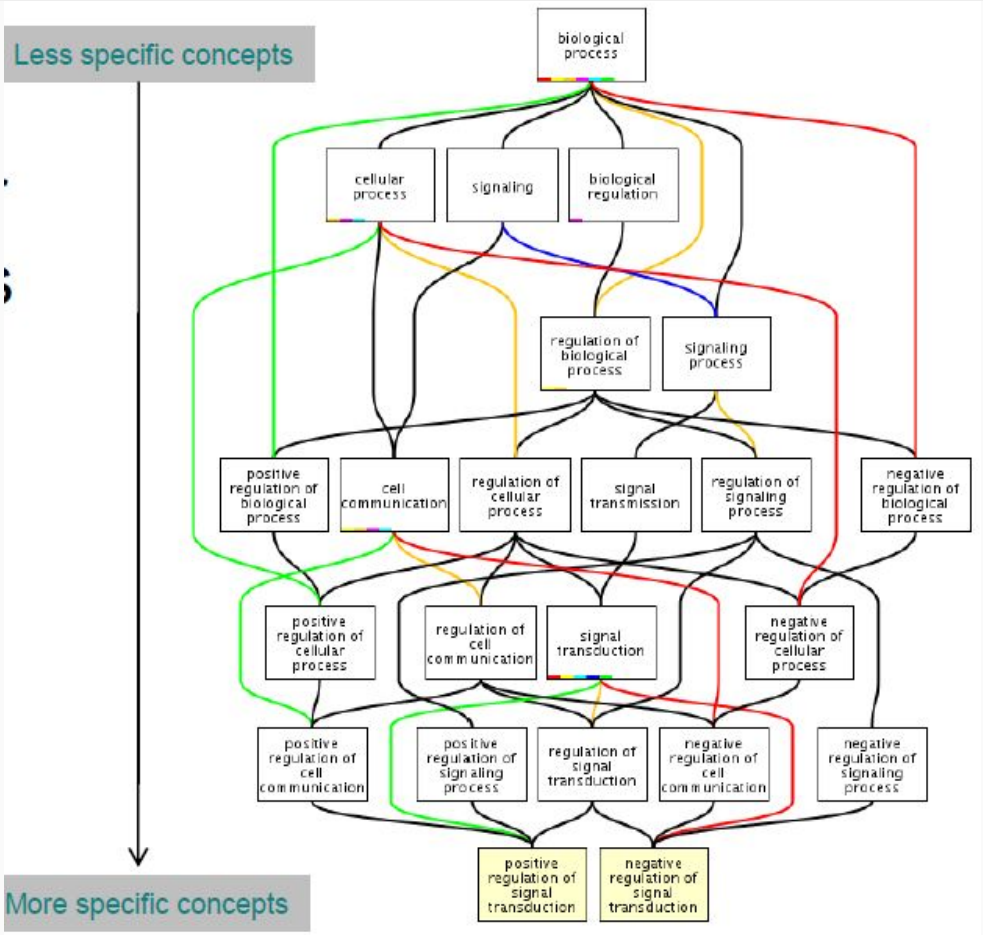
Hallmarks of Cancer genesets

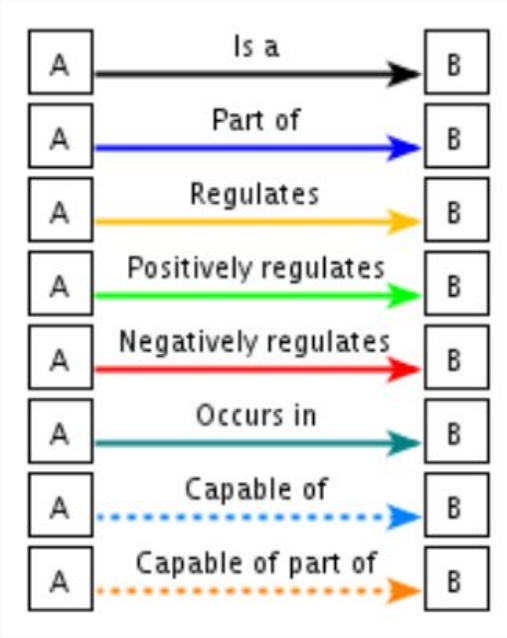


Gene Ontology (GO)

- Ontology: "*a set of concepts and categories in a subject area or domain that shows their properties and the relations between them*"
- Gene Ontology: a controlled vocabulary of biological properties and their relationships
- GO composed of **terms** organized into three **namespaces**

Gene Ontology





GO Term Namespaces

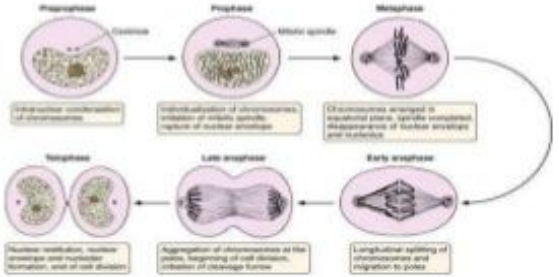
1. Molecular Function

An elemental activity or task or job

- protein kinase activity
- insulin receptor activity

2. Biological Process

A commonly recognized series of events



- cell division

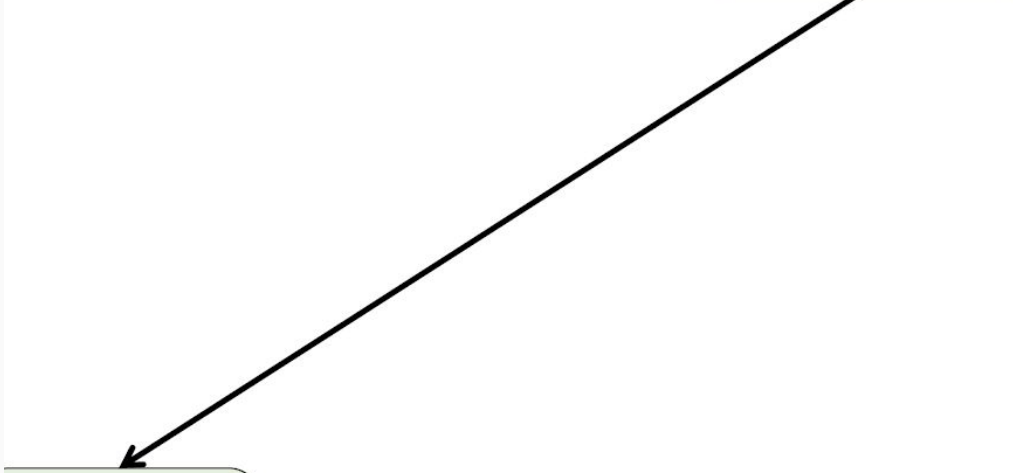
3. Cellular Component

Where a gene product is located

- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

A GO annotation is …

…a statement that a gene product;

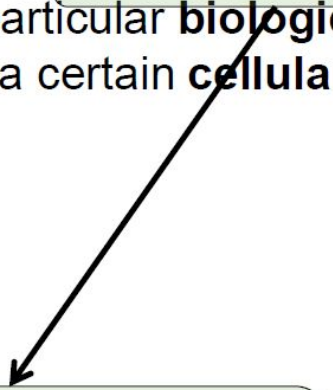


Accession	Name	GO ID	GO term name	Reference	Evidence code
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
   or is involved in a particular **biological process**
   or is located within a certain **cellular component**



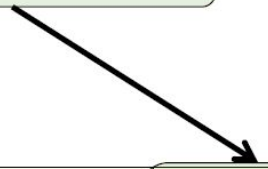
Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
   or is involved in a particular **biological process**
   or is located within a certain **cellular component**

2. as described in a particular reference



Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

A GO annotation is …

…a statement that a gene product;

1. has a particular molecular function
   or is involved in a particular biological process
   or is located within a certain cellular component

2. as described in a particular reference

3. as determined by a particular method

Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

GO can add biological meaning to your data !!

BMC Endocr Disord. 2013 Oct 7;13(1):43. [Epub ahead of print]

Gene expression of sternohyoid and diaphragm muscles in type 2 diabetic rats.

van Lunteren E, Moyer M.

Abstract

BACKGROUND: Type 2 diabetes differs from type 1 diabetes in its pathogenesis. Type 1 diabetic diaphragm has altered gene expression which includes lipid and carbohydrate metabolism, ubiquitination and oxidoreductase activity. The objectives of the present study were to assess respiratory muscle gene expression changes in type 2 diabetes and to determine whether they are greater for the diaphragm than an upper airway muscle.

METHODS: Diaphragm and sternohyoid muscle from Zucker diabetic fatty (ZDF) rats were analyzed with Affymetrix gene expression arrays.

RESULTS: The two muscles had 97 and 102 genes, respectively, with at least ± 1.5-fold significantly changed expression with diabetes, and these were assigned to gene ontology groups based on over-representation analysis. Several significantly changed groups were common to both muscles, including lipid metabolism, carbohydrate metabolism, muscle contraction, ion transport and collagen, although the number of genes and the specific genes involved differed considerably for the two muscles. In both muscles there was a shift in metabolism gene expression from carbohydrate metabolism toward lipid metabolism, but the shift was greater and involved more genes in diabetic diaphragm than diabetic sternohyoid muscle. Groups present in only diaphragm were blood circulation and oxidoreductase activity. Groups present in only sternohyoid were immune and inflammation and response to stress and wounding, with complement genes being a prominent component.

World J Gastroenterol. 2013 Jun 7;19(21):3249-54. doi: 10.3748/wjg.v19.i21.3249.

Early dynamic transcriptomic changes during preoperative radiotherapy in patients with rectal cancer: a feasibility study.

Supiot S, Gouraud W, Campion L, Jezéquel P, Buecher B, Charrier J, Heymann MF, Mahé MA, Rio E, Chérel M.

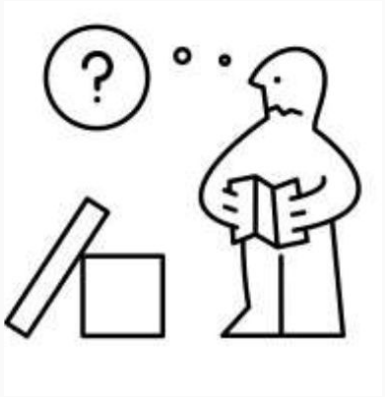
Abstract

AIM: To develop novel biomarkers of rectal radiotherapy, we measured gene expression profiles on biopsies taken before and during preoperative radiotherapy.

RESULTS: Microarray analysis showed that preoperative radiotherapy significantly up-regulated 31 genes and down-regulated 6 genes. According to the Gene Ontology project classification, these genes are involved in protein metabolism (ADAMDEC1; AKAP7; CAPN5; CLIC5; CPE; CREB3L1; NEDD4L; RAB27A), ion transport (AKAP7; ATP2A3; CCL28; CLIC5; F2RL2; NEDD4L; SLC6A8), transcription (AKAP7; CREB3L1; ISX; PABPC1L; TXNIP), signal transduction (CAPN5; F2RL2; RAB27A; TNFRSF11A), cell adhesion (ADAMDEC1; PXDN; SPON1; S100A2), immune response (CCL28; PXDN; TNFRSF11A) and apoptosis (ITM2C; PDCD4; PVT1). Up-regulation of 3 genes (CCL28; CLIC5; PDCD4) was detected by 2 different probes and up-regulation of 2 genes (RAB27A; TXNIP) by 3 probes.

Challenges using GO

- Enriched categories are often too general to be useful:
  - e.g. GO:0003700 - DNA-binding transcription factor activity
- Hierarchical nature of GO creates redundant results
  - GO:0003700 - DNA-binding transcription factor activity “IS A”
  - GO:0006355 - regulation of transcription, DNA-templated



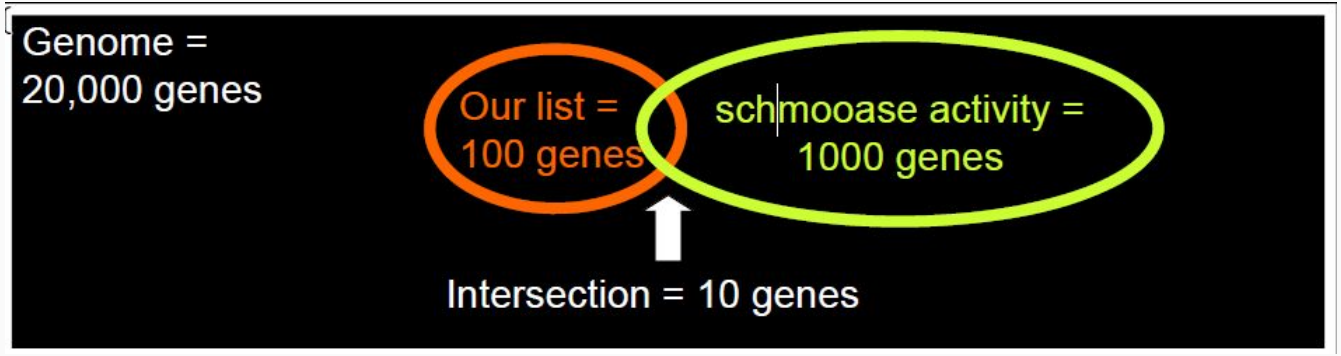
Enrichment Methods and Tools

“Enrichment” and Geneset

- Enrichment “act of making fuller or meaningful” - (dictionary.com)
- Gene Set Enrichment:
  Does our gene list overlap a given gene set more than we would expect by chance?
- Geneset are enriched if experimental findings are in accordance with set of interest

Why do enrichment analysis?

- "See the forest for the trees": High-level biological picture not visible when considering genes one-by-one
- Too many genes to examine in detail
  - Strong perturbations may yield 1000s of DE genes
  - How do we know that what we're seeing is surprising?



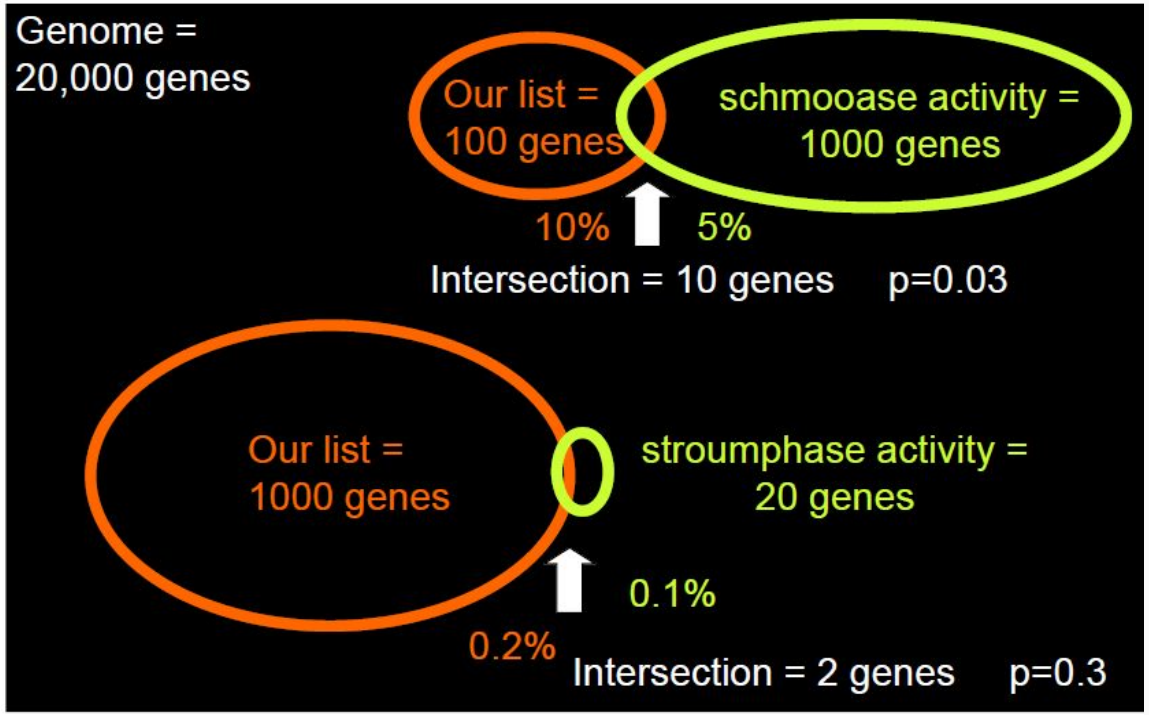
Gene Set Enrichment Analysis

- Gene set enrichment analysis:
  - Compute a statistic that indicates the degree of similarity between a gene list and a gene set
  - Statistic compared to a “null” or “background” distribution to calculate a p-value
- Statistic may indicate gene list:
  - Has larger overlap with gene set than expected
  - Has smaller overlap with gene set than expected
  - Has overlapping genes that are more highly or lowly ranked than expected

Main Types of Enrichment Analysis

- List-based: inputs are
  - A subset of all genes chosen by some relevant method
  - A list of annotations, each linked to genes
- Rank-based: inputs are
  - A set of all genes ranked by some metric (ratio, foldchange, etc.)
  - A list of annotations, each linked to genes
- List-based with relationships: inputs are
  - A subset of all genes
  - A list of annotations, each linked to genes, organized in some relationship (e.g., a hierarchy)

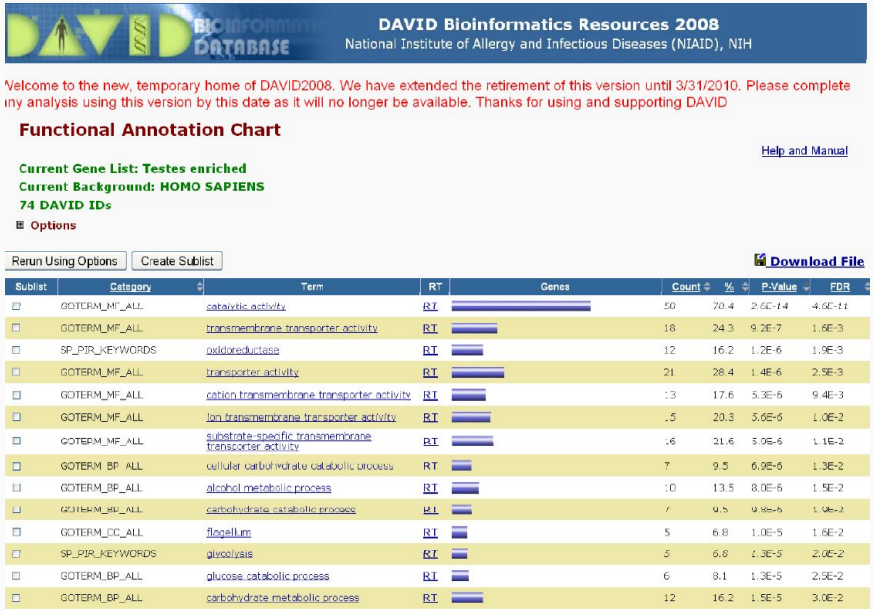
Statistic to test for enrichment



- Test for enrichment
  - **Fisher's exact**
  - **Hypergeometric**
  - Binomial
  - Chi-squared
  - Kolmogorov-Smirnov
  - Permutation

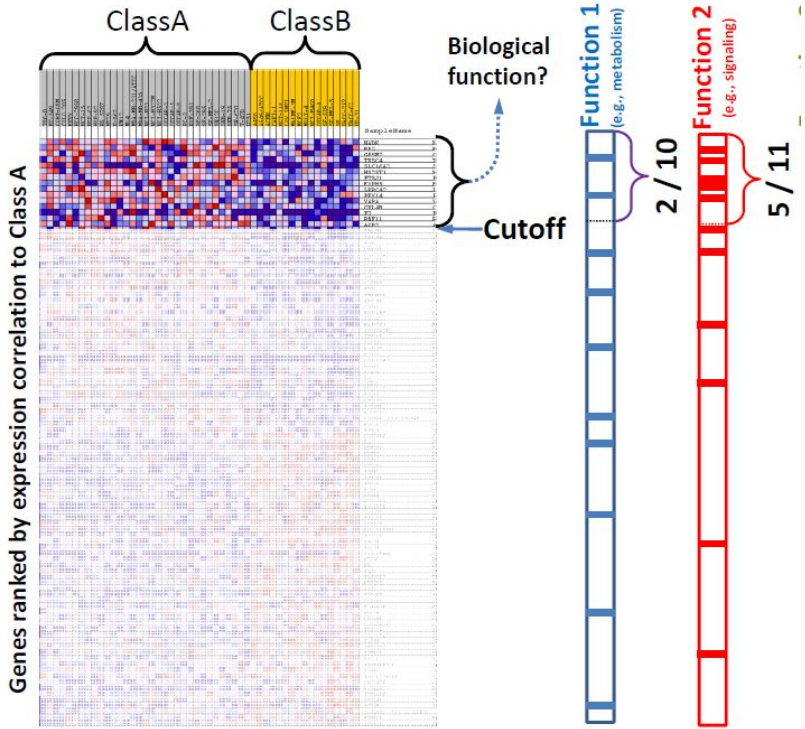
DAVID

- Database for Annotation, Visualization and Integrated Discovery (NIAID) http://david.abcc.ncifcrf.gov/
- List-based; Lots of identifiers; lots of species
- Allows background definition
- Statistic is a modified Fisher exact test



Overrepresentation vs Aggregate score

- So far, methods use intuition of “over-representation”
- Genes of interest can be defined in many different ways:
  - By p-value (which to use?)
  - Filtered by fold change?
  - Ignores potentially meaningful information on other genes
- Aggregate score methods use statistics for all genes in a dataset



Gene Set Enrichment Analysis (GSEA)

- Choosing gene list thresholds is hard and subjective
- Expression datasets measure thousands of genes at once
- Ideally use all information from an expression
- Gene Set Enrichment Analysis: statistical method that uses ranked gene expression statistics to calculate enrichment

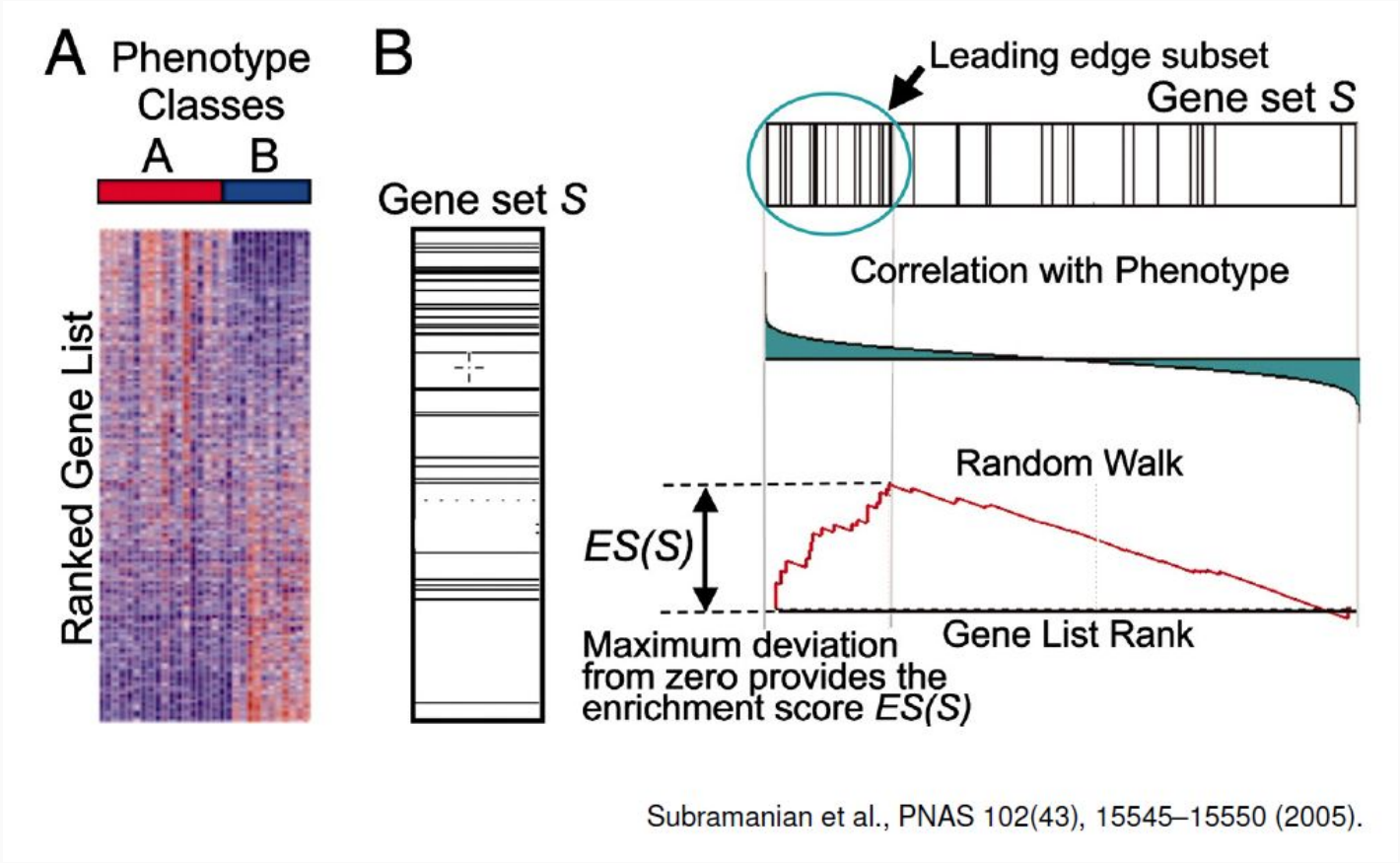
Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander and Jill P. Mesirov

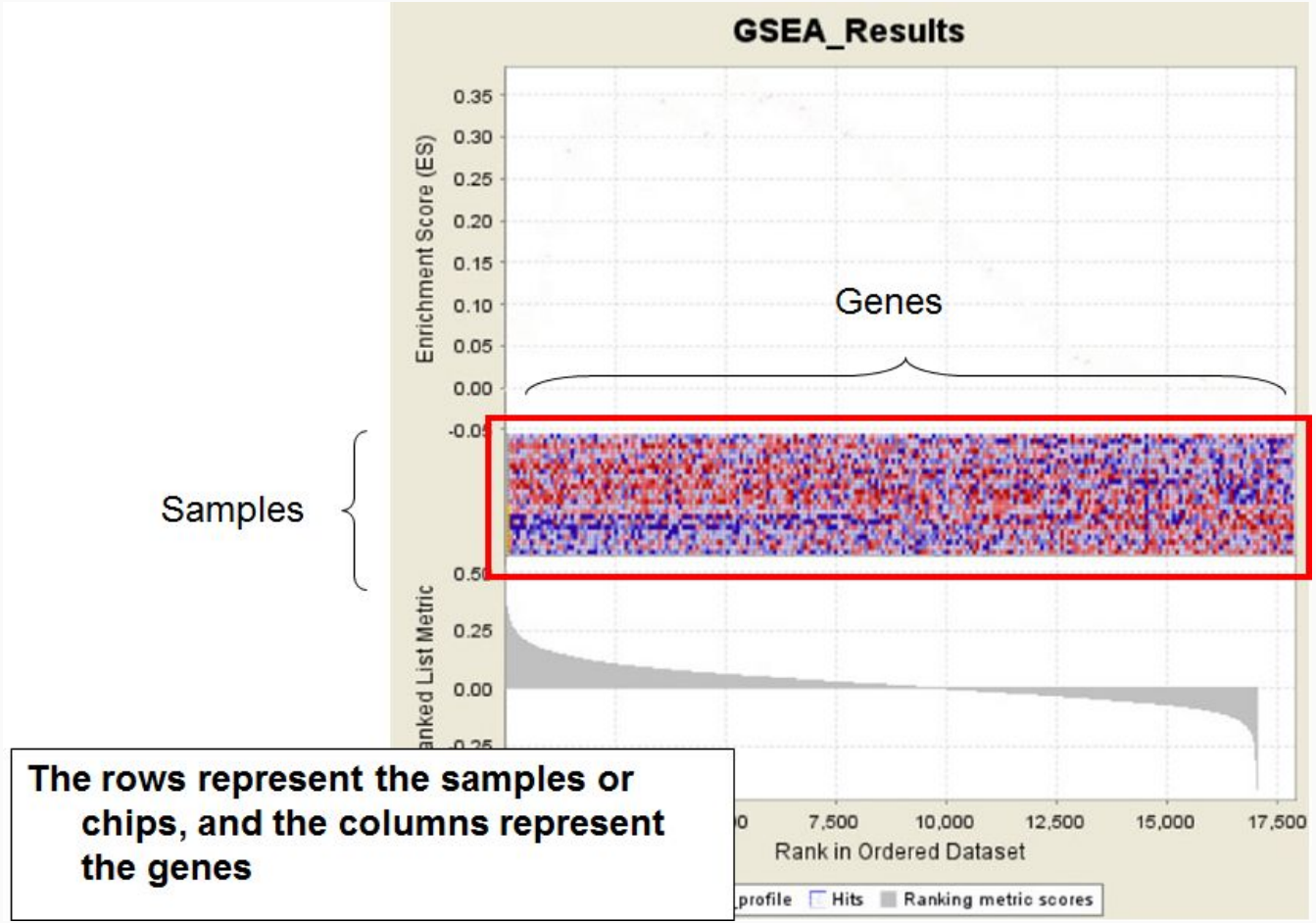
PNAS 2005 October, 102 (43) 15545-15550. https://doi.org/10.1073/pnas.0506580102

Contributed by Eric S. Lander, August 2, 2005

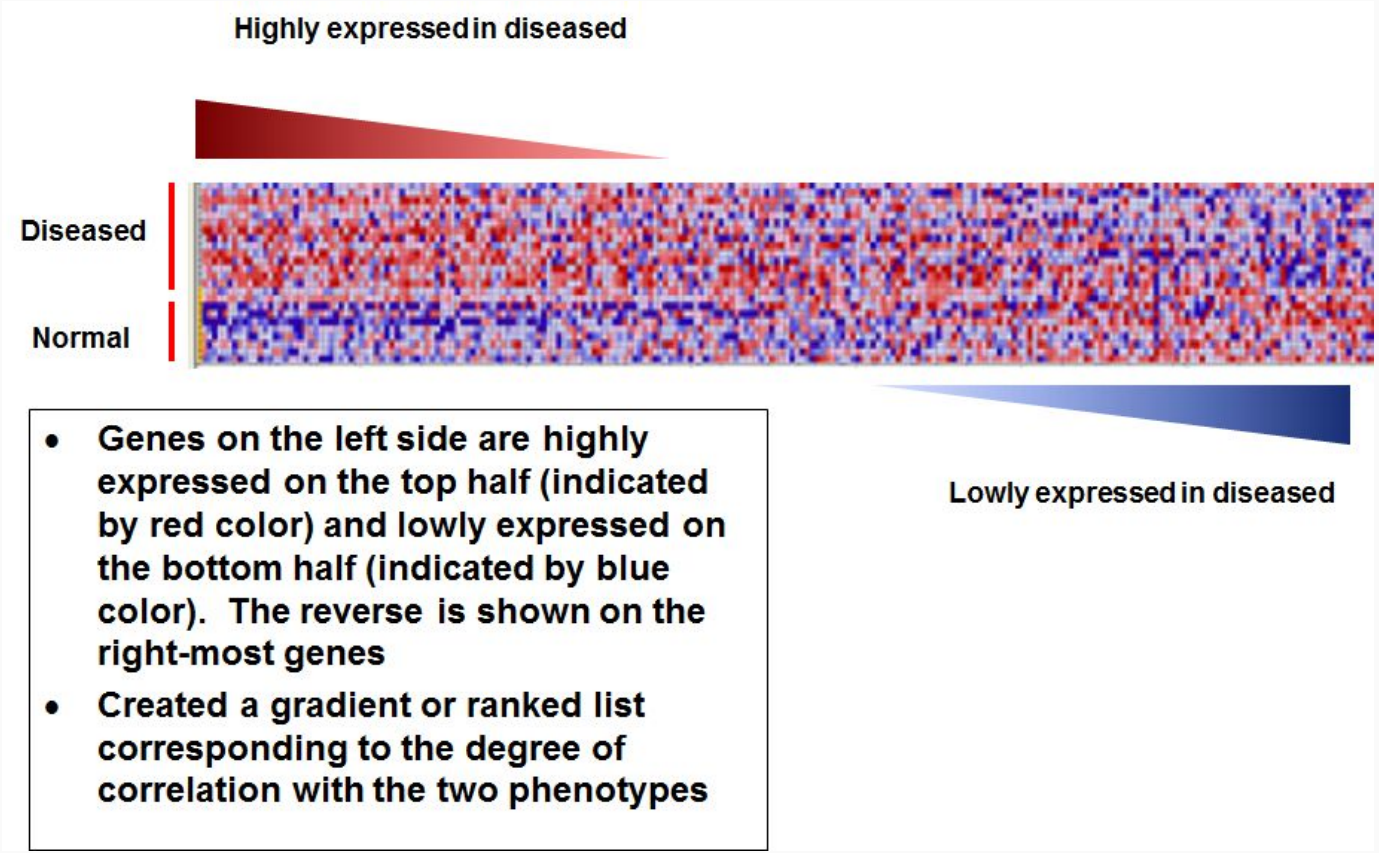
Schematic Overview of GSEA



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

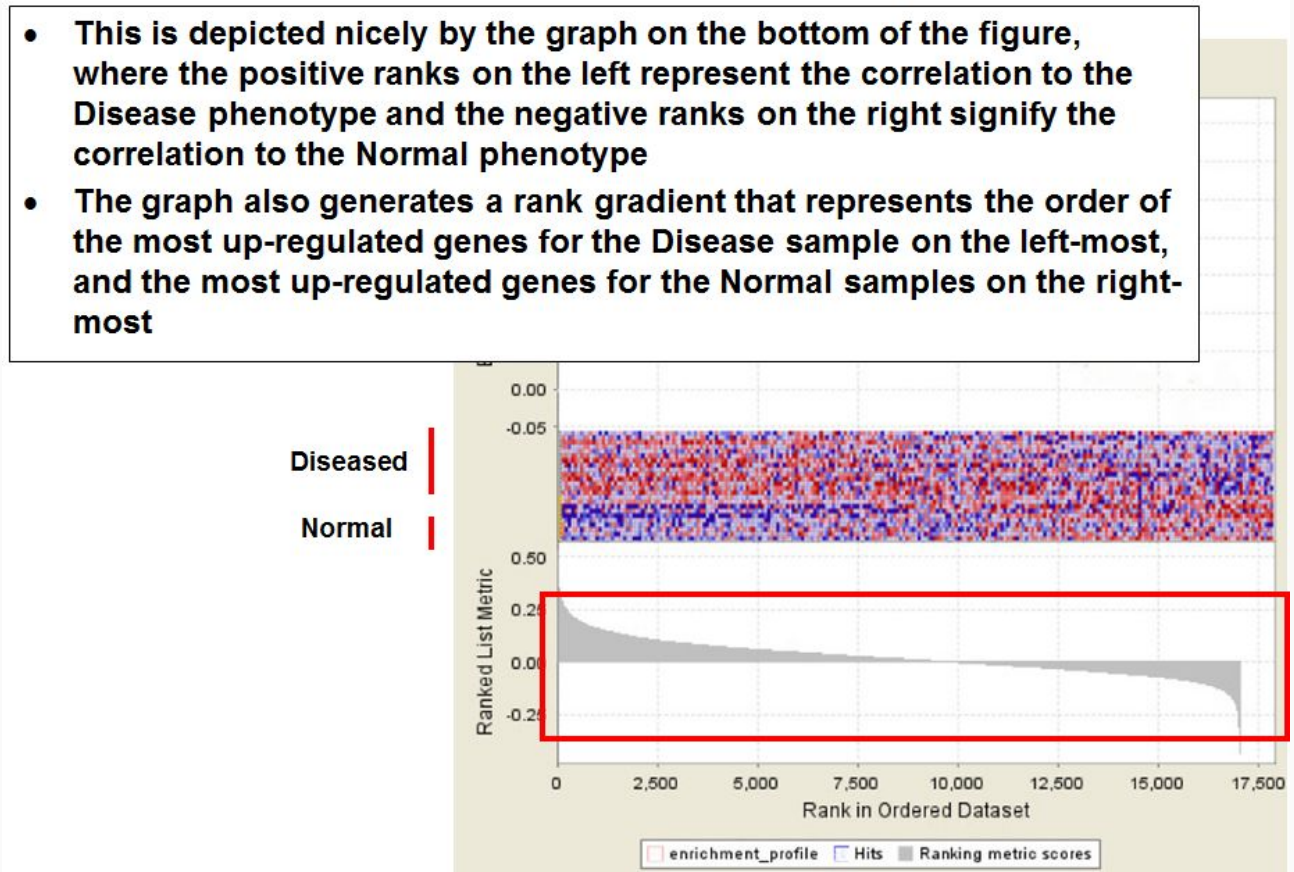


The rows represent the samples or chips, and the columns represent the genes

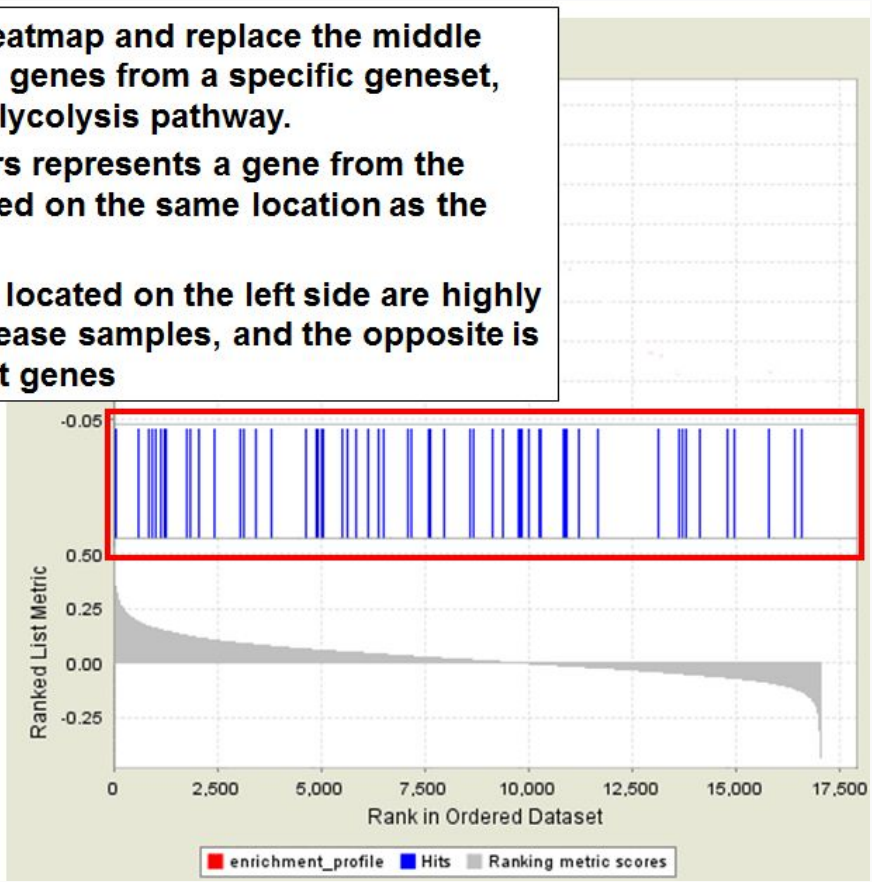


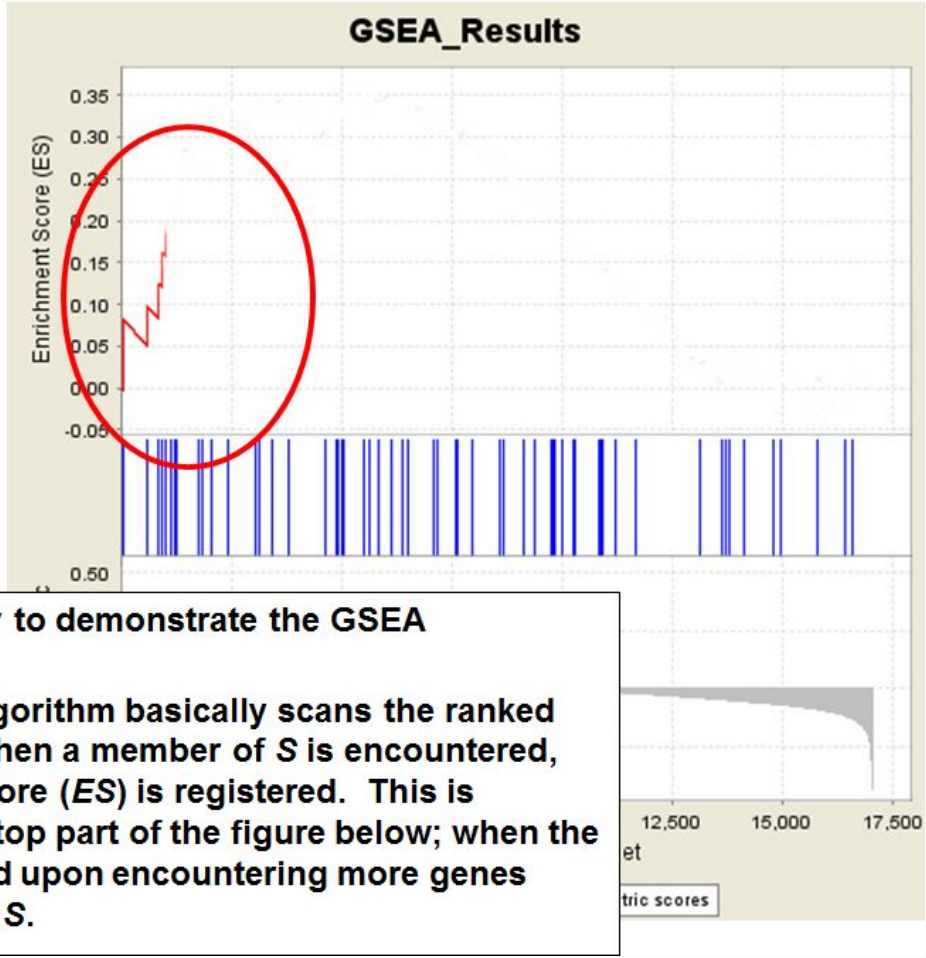
- Genes on the left side are highly expressed on the top half (indicated by red color) and lowly expressed on the bottom half (indicated by blue color). The reverse is shown on the right-most genes
- Created a gradient or ranked list corresponding to the degree of correlation with the two phenotypes

- **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**

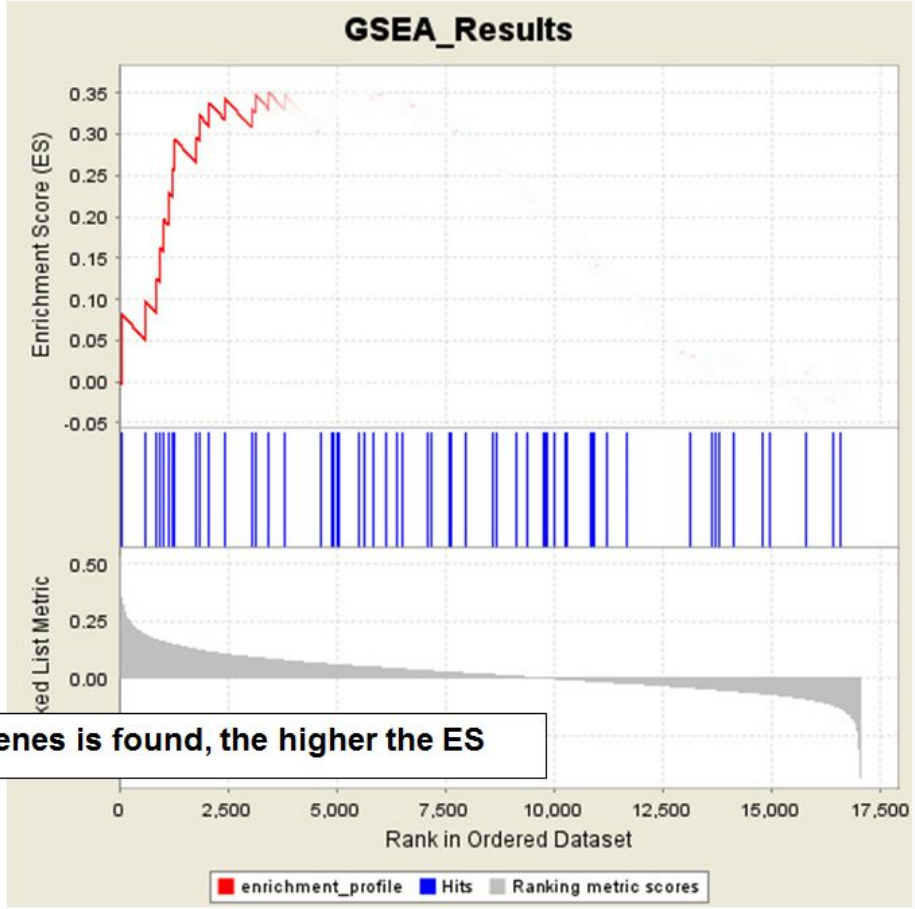


- Now, let's hide the heatmap and replace the middle part of the figure with genes from a specific geneset, say genes from the Glycolysis pathway.
- Each vertical blue bars represents a gene from the pathway, being mapped on the same location as the whole dataset
- Again, genes that are located on the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

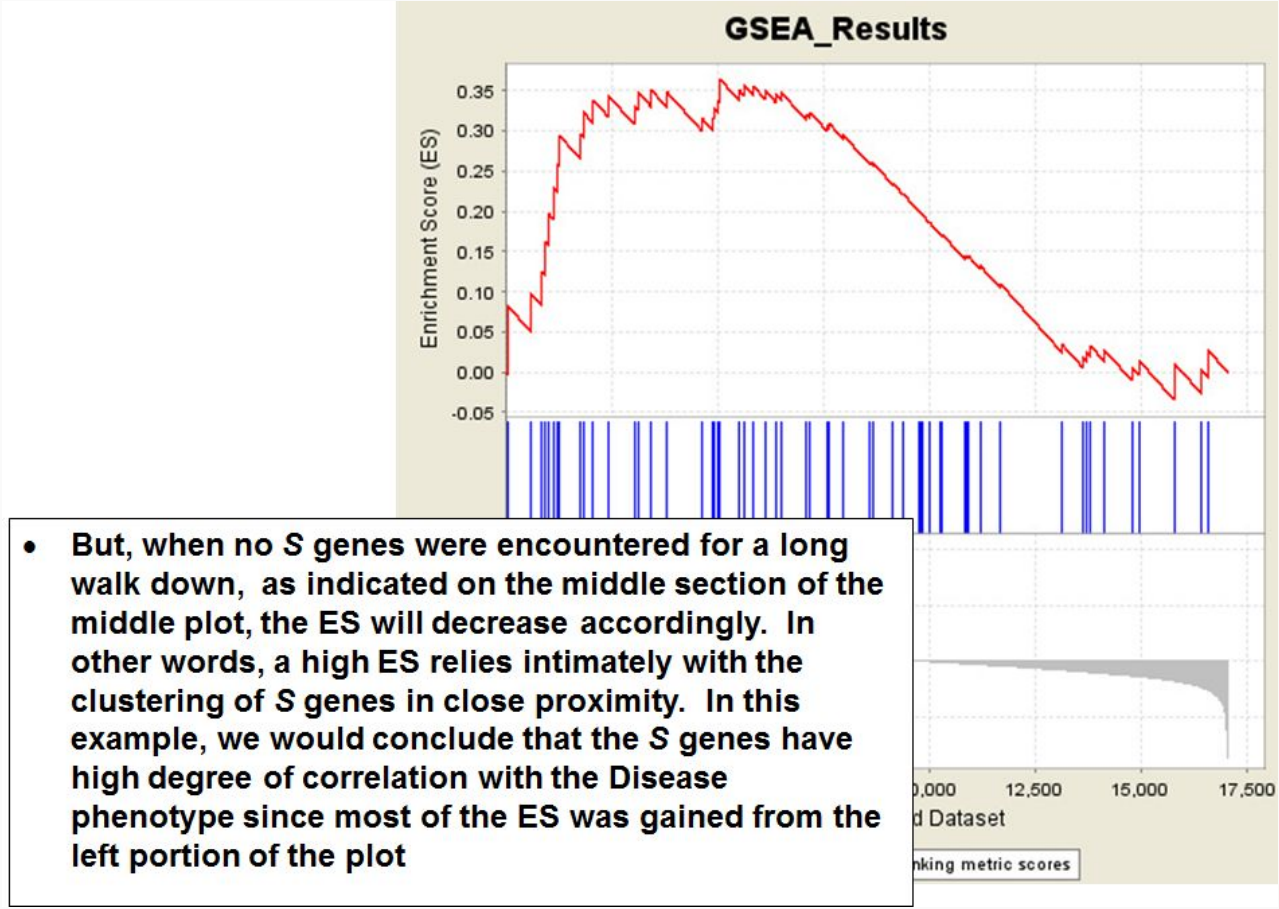




- **Now, we are ready to demonstrate the GSEA algorithm.**
- **The walk down algorithm basically scans the ranked gene list *L*, and when a member of *S* is encountered, an Enrichment Score (*ES*) is registered. This is illustrated on the top part of the figure below; when the ES started to build upon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



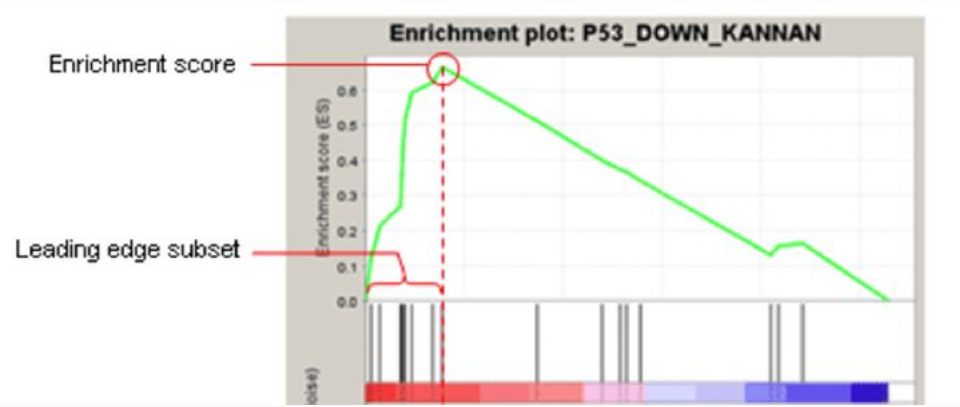
- The more *S* genes is found, the higher the ES



- 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

Leading Edge Genes

- Leading edge subset of a gene set = the genes that appear in the ranked list before the running sum reaches the max value.



- Leading edge analysis = examine the genes that are in the leading edge subsets of the enriched gene sets.
- For a negative ES, it is the set of members that appear subsequent to the peak score.

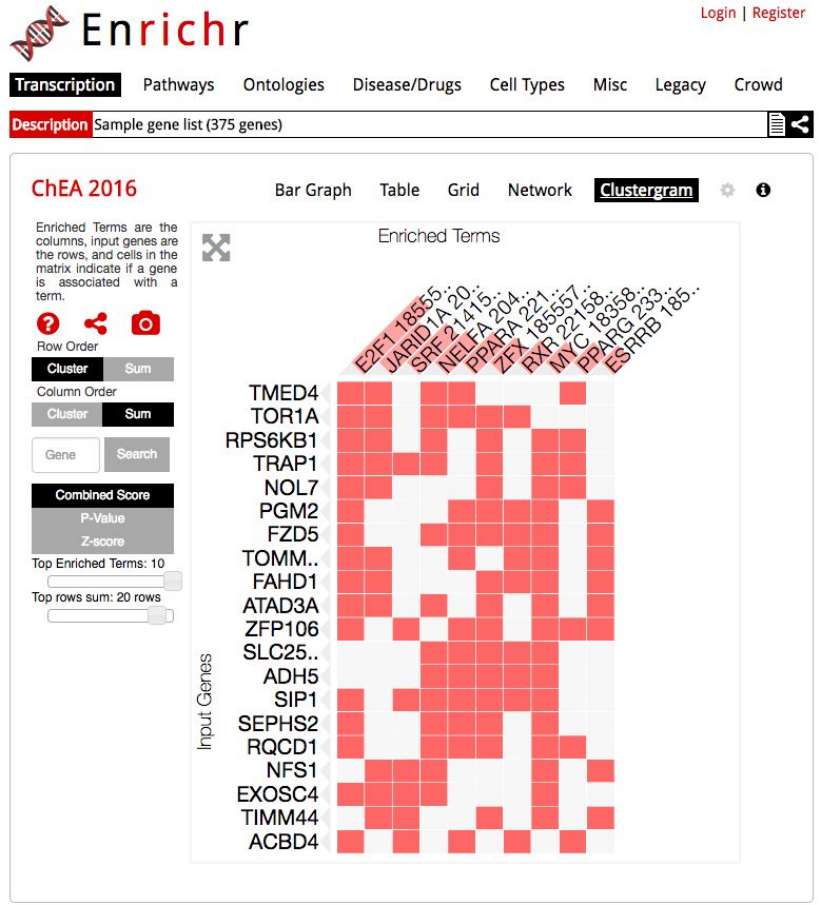
GSEA Statistic

- Enrichment score (ES) is *directional*:
  - positive ES: gene in geneset at the top of the ranked list
  - negative ES: gene in geneset at the bottom of the ranked list
- Normalized Enrichment score (**NES**): ES adjusted for gene set size
  - can be used to compare analysis results across gene sets
- Nominal p-value estimates the statistical significance of the enrichment score for a single gene set
- Must correct for multiple hypothesis testing
  - Bonferroni
  - **Benjamini-Hochberg (a.k.a. FDR)**
- False Discovery Rate (FDR) is the estimated fraction of results expected to be false positives
  - e.g. for 100 significant results at FDR < 0.05, 5 are likely to be FP

Advantages of GSEA

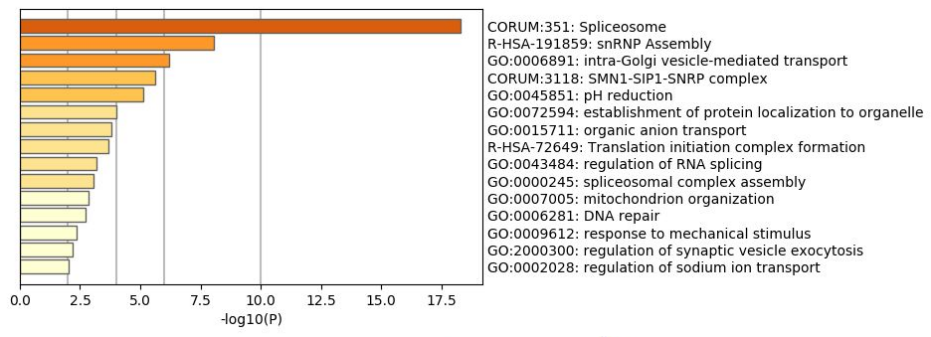
- Agnostic to the type of gene set and the source of annotation
- Operates on any ordered gene list
- Does not require the choice of a gene selection threshold or the explicit definition of a statistically significant marker set
- Uses distribution-free, non-parametric, permutation-based test procedures with increased statistical power
- Incorporates the permutation of phenotype labels thereby preserving the “biological” correlation structure of the markers
- Takes into account multiple hypotheses testing of multiple gene sets.

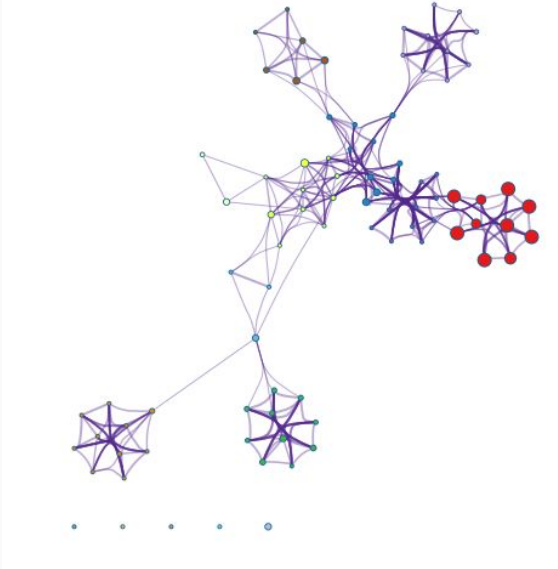
Enrichr



- The enrichment analysis tool http://amp.pharm.mssm.edu/Enrichr/
- Clustergram to produce dynamic heatmaps of enriched terms as columns and user input genes as rows
- helps understand the relationships between their input genes and enriched terms.

Metascape





- metascape.org
- Maintains set of curated geneset databases (GO, KEGG, etc)
- Computes hyperenrichment
- Clusters gene sets by shared gene statistic
- Easy to use
- Makes nice figures

Considerations and Recommendations

- **Choose a tool that**
  - Includes your species
  - Includes your gene / probe identifiers
  - Has up-to-date annotation
  - Lets you define your background (if possible)
- **Try at least a few tools, inspect for agreement**
- **Try different list definitions** (e.g. different p-value, log fold change thresholds)