



# Enrichment Analysis

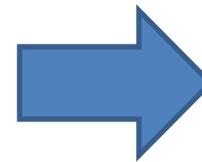
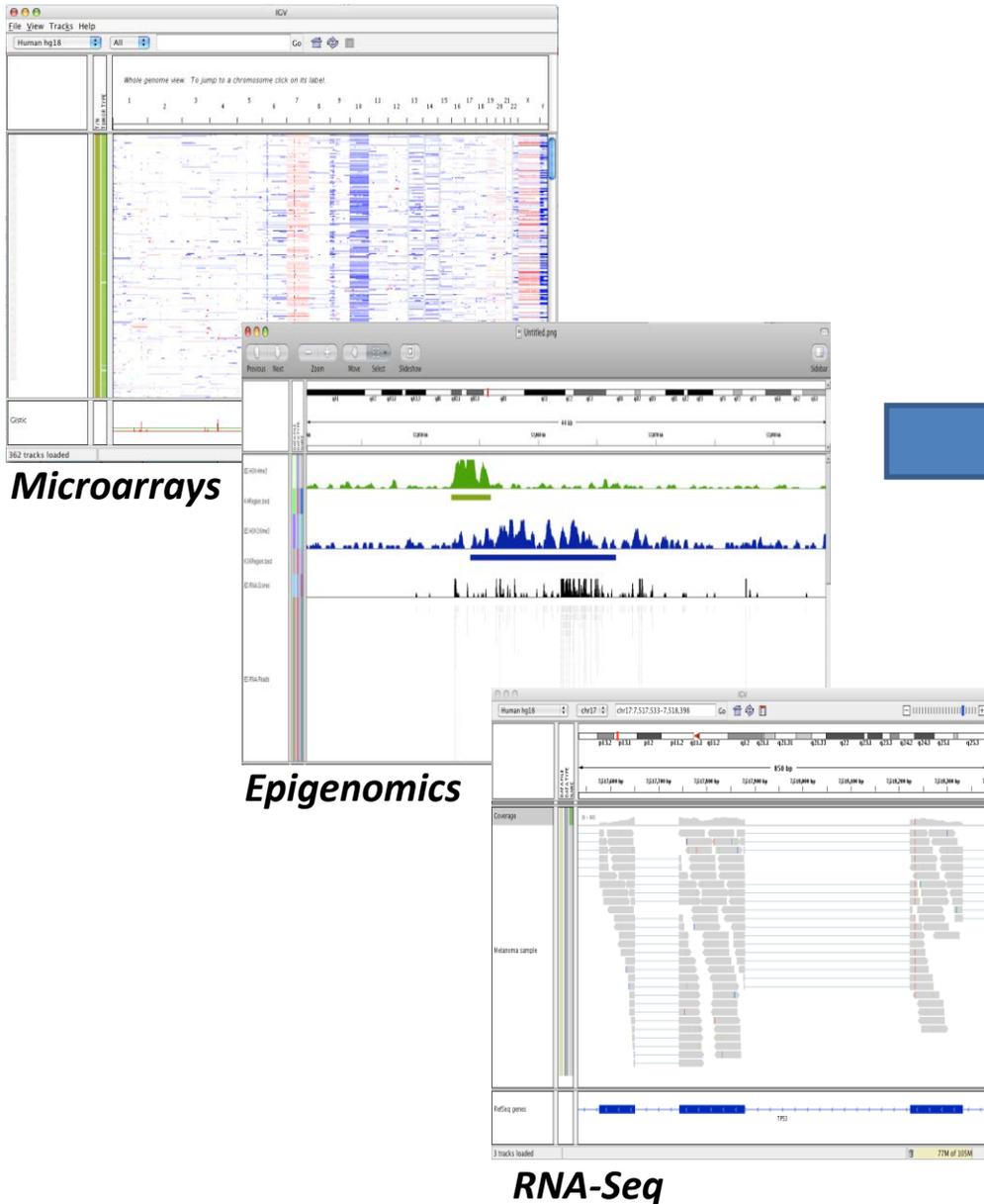
Prat Thiru



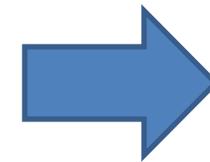
# Outline

- Overview and Goals of Enrichment Analysis
- Databases for Gene Set Annotations
- Statistics for Enrichment
- Enrichment Tools
- Practicalities
- Supplementary Information on Statistics

# Overview



Gene	Log Ratio	p-value
Abcg1	-2.09614	4.72E-07
Adamts5	2.483321	1.33E-07
Alox12b	-2.41347	3.59E-07
Arg1	-2.27214	3.06E-07
AU018091	2.048711	4.62E-07
Bex1	2.591349	4.08E-07
Degs2	-2.46253	1.54E-07
Klk7	-2.18902	3.77E-07
Krt78	-2.89916	2.18E-07
Ly6c1	3.085592	9.41E-08
Ly6g6c	-2.55108	3.62E-07
Sdr16c6	-2.16277	4.05E-07
Sdr9c7	-2.25984	2.63E-07
Sept5	-2.08797	6.31E-07
Kprp	-2.34542	6.77E-07
Ly6a	2.839925	6.04E-07
Slc2a3	2.199118	6.52E-07
Sprr2i	-2.22872	5.67E-07
Mxd1	-1.77522	9.66E-07
Cidea	-1.93749	1.20E-06
Krt16	-1.91642	1.24E-06
Krt8	2.057569	1.22E-06
Trex2	-1.71243	1.29E-06
Aldh3b2	-1.7556	2.63E-06
Asprv1	-1.56796	2.35E-06



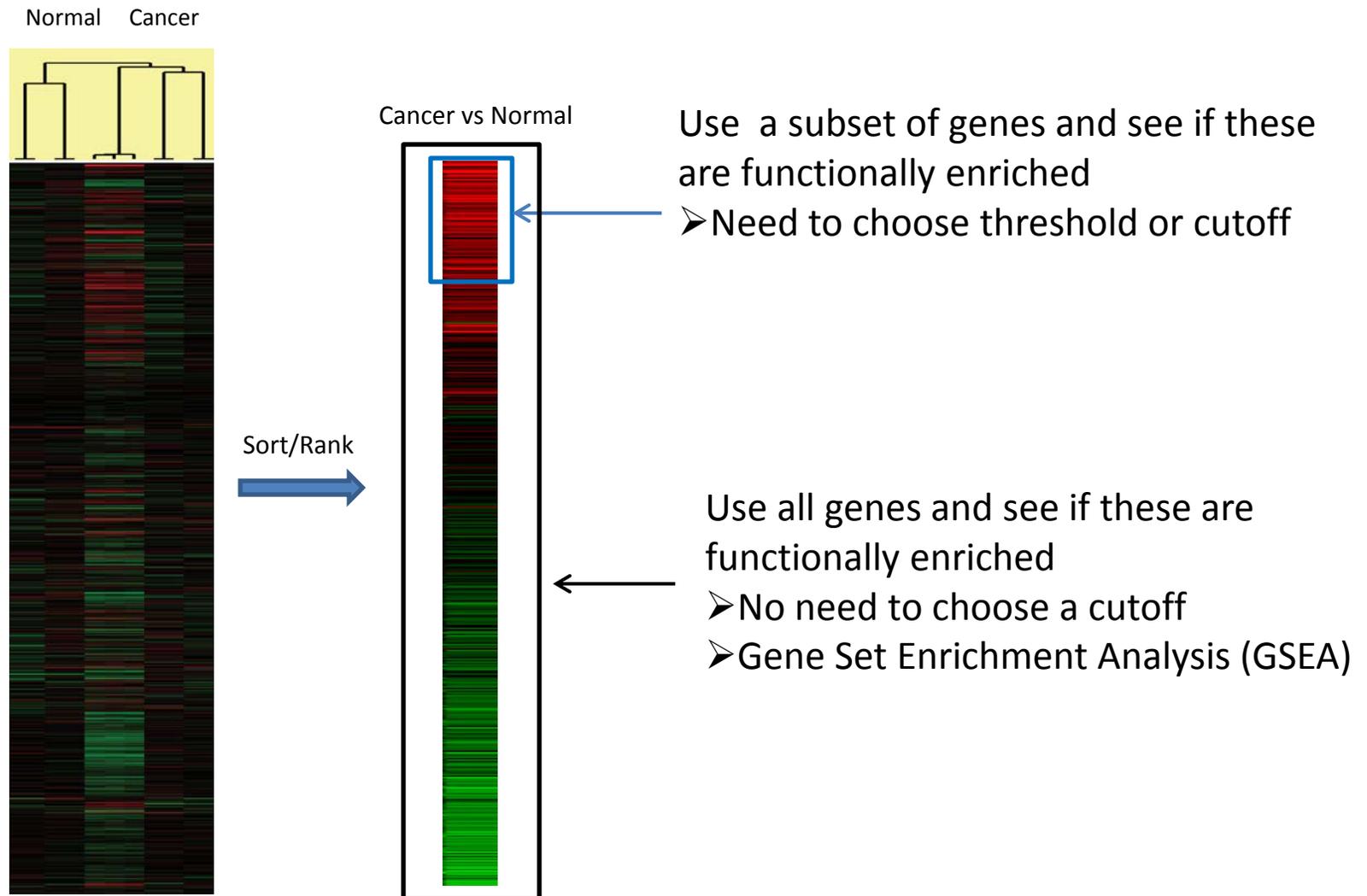
?

Long List of Genes

# Goals of Enrichment Analysis

- Identifying the differences in a set of genes will give more biological insight than an individual gene
- Functional annotations that are over-represented in the gene list
- Find related genes, for eg. by metabolic pathways, cell signaling pathways, type of kinase, targets of miRNA, etc.

# Enrichment Analysis: Two Strategies



# Databases for Annotations

Database	Description	Website
KEGG	Metabolic Pathways	<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>
Gene Ontology (GO)	Controlled vocabulary for genes (and gene products)	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
MSigDB	Molecular signatures database: a collection of annotated gene sets	<a href="http://www.broadinstitute.org/gsea/msigdb/index.jsp">http://www.broadinstitute.org/gsea/msigdb/index.jsp</a>
DAVID*	Various annotations: Panther, Pfam, COG and more	<a href="http://david.abcc.ncifcrf.gov">http://david.abcc.ncifcrf.gov</a>

\*a collection of databases

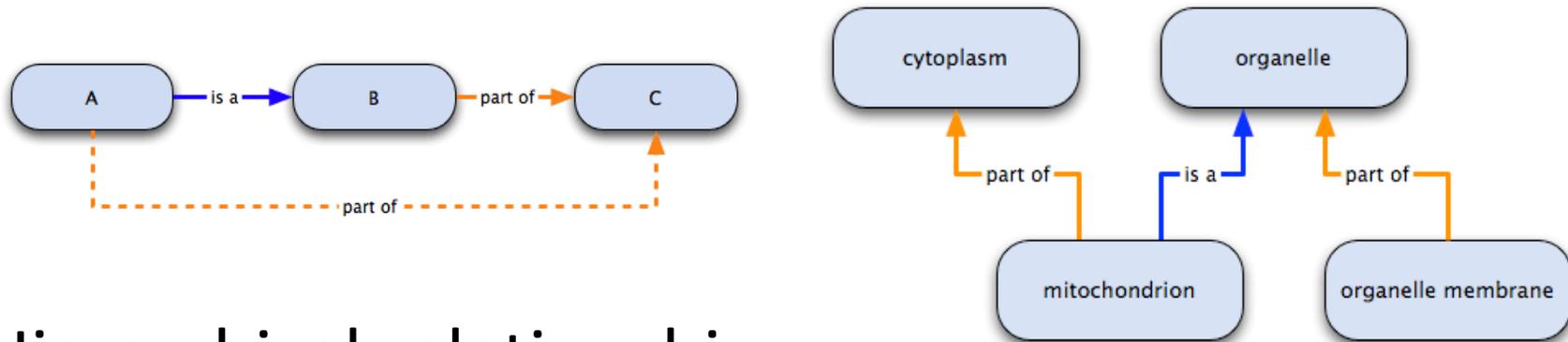
- Other custom or user-defined gene sets
  - different stages of development (eg. erythropoiesis)

# Gene Ontology (GO)

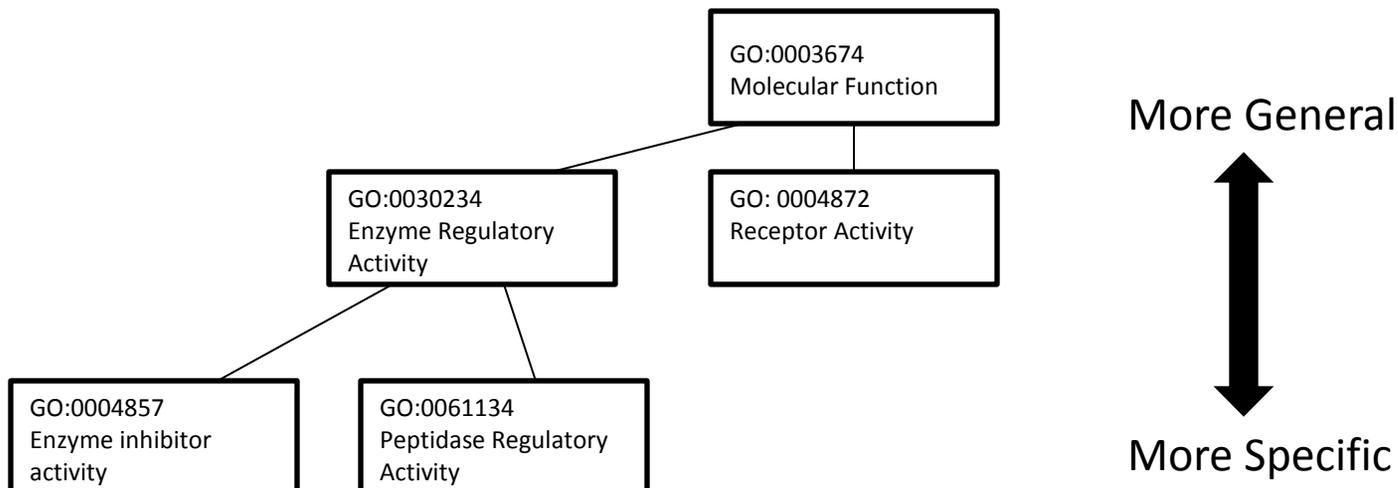
- Collection of gene sets with controlled vocabulary
- Cellular Component: parts of a cell (eg. nucleus, ER)
- Molecular Function: activity of a gene product (eg. binding, catalysis)
- Biological Process: series of events accomplished by one or more ordered assemblies of molecular functions (eg. pyrimidine metabolic process)
  - Function vs Process: the process must have more than one distinct steps.

# Gene Ontology

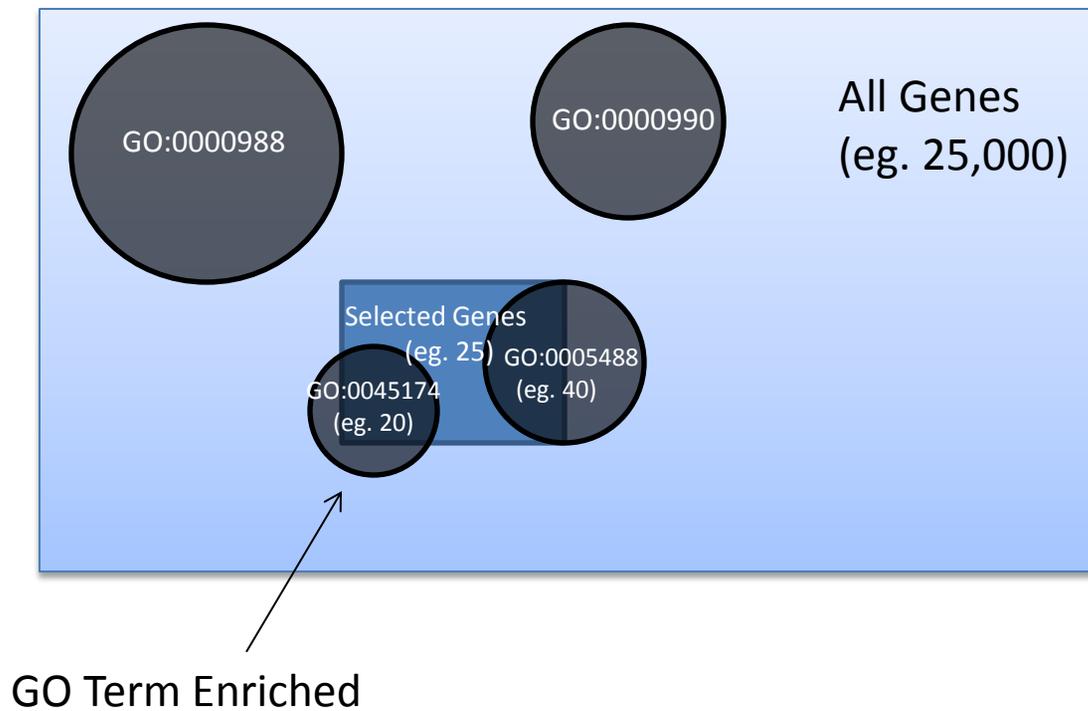
- Based on “is a” or “part of” relationship



- Hierarchical relationship

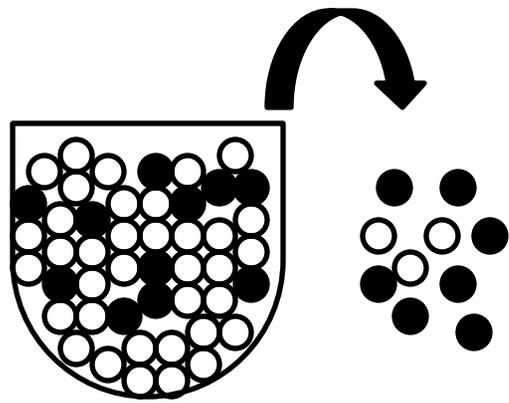


# GO Enrichment



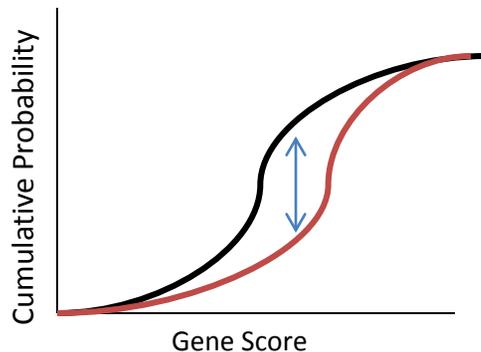
# Assessing Significance of Enrichment

## ➤ Fisher's Exact Test (Hypergeometric Test)



What is the probability of getting 7 or more black balls?

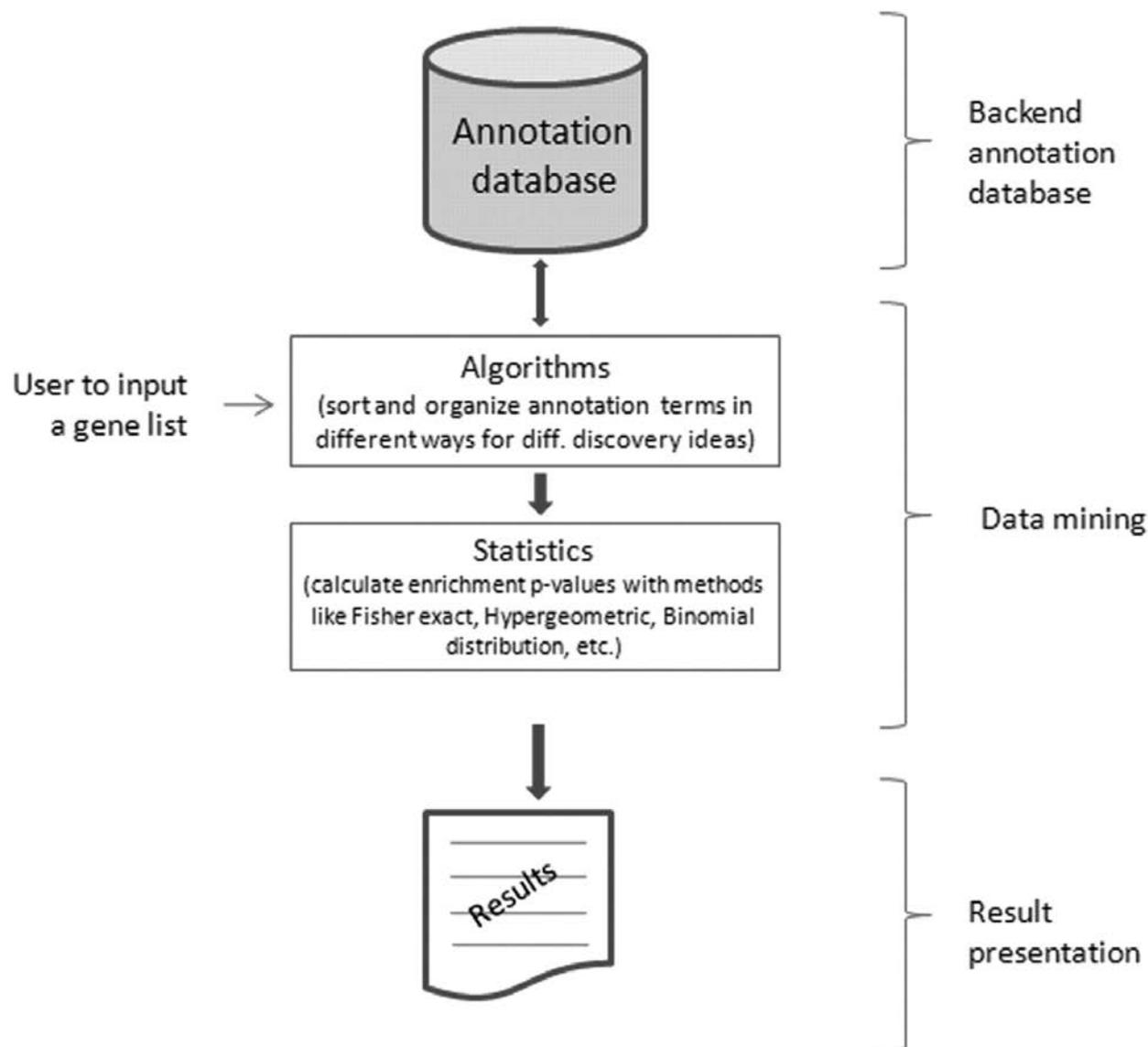
## ➤ Kolmogorov-Smirnov (KS) Test



# Assessing Significance of Enrichment: Which test?

- One vs two-sided tests
  - Testing only for enrichment vs enrichment/depletion
- Sample size
  - Size of sample (eg. small vs large) important when choosing a test
- p-values
  - Useful for ranking
  - Dependent on the test
- Corrected p-values
  - p-value should be corrected because of multiple hypothesis testing

# Enrichment Analysis Tool Infrastructure



# Enrichment Analysis Tools

Tool	Statistical Method	Website
DAVID	Fisher	<a href="http://david.abcc.ncifcrf.gov">http://david.abcc.ncifcrf.gov</a>
GSEA	KS Test	<a href="http://www.broadinstitute.org/gsea">http://www.broadinstitute.org/gsea</a>
BiNGO (Cytoscape Plugin)	Hypergeometric; Binomial	<a href="http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html">http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html</a>
GeneGO*	Hypergeometric	<a href="http://www.genego.com">http://www.genego.com</a>
GoMiner	Fisher	<a href="http://discover.nci.nih.gov/gominer/index.jsp">http://discover.nci.nih.gov/gominer/index.jsp</a>

# Enrichment Analysis: Factors to Consider

- Gene list
- Background gene list
- Statistical test
- Gene set annotations (including user-defined)
- p-value Correction



# DAVID

**DAVID: Functional Annotation Tools - Mozilla Firefox**

File Edit View History Bookmarks Tools Help

DAVID: Functional Annotation Tools

david.abcc.ncifcrf.gov/tools.jsp

Most Visited BaRC: Bioinformatics ... BLAST Help Tak 2.0 BaRC Wiki EMBOSS WIBR String counter NCBI Perl mod UCSC Genome Browse... BaRC analysis to

**Analysis Wizard**  
DAVID Bioinformatics Resources 6.7, NIAID/NIH

Home Start Analysis Shortcut to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us

**Upload** List Background

Upload Gene List

[Demolist 1](#) [Demolist 2](#)

[Upload Help](#)

Step 1: Enter Gene List

A: Paste a list

QPRT  
GALNT3  
PRDX2  
GCH1

Clear

Or

B: Choose From a File

Browse...

Multi-List File ?

Step 2: Select Identifier

OFFICIAL\_GENE\_SYMBOL

Step 3: List Type

Gene List

Background

Step 4: Submit List

Submit List

[Tell us how you like the tool](#)

[Contact us for questions](#)

← Step 1. Submit your gene list through left panel.

An example:

Copy/paste IDs to "box A" -> Select Identifier as "Affy\_ID" -> List Type as "Gene List" -> Click "Submit" button

1007\_s\_at  
1053\_at  
117\_at  
121\_at  
1255\_g\_at  
1294\_at  
1316\_at  
1320\_at  
1405\_i\_at  
1431\_at  
1438\_at  
1487\_at  
1494\_f\_at  
1598\_g\_at

Gene List

Choose Identifier

Select List Type



# DAVID: Output

**Annotation Summary Results**

Current Gene List: List\_1  
 Current Background: Homo sapiens  
 22793 DAVID IDs  
 Check Defaults  Clear All

**Functional Annotation Chart**

Current Gene List: List\_1  
 Current Background: Homo sapiens  
 22793 DAVID IDs

**Options**

Thresholds: Count  EASE

Display:  Fold Enrichment  Bonferroni  Benjamini  FDR  Fisher Exact  LT,PH,PT # of Records

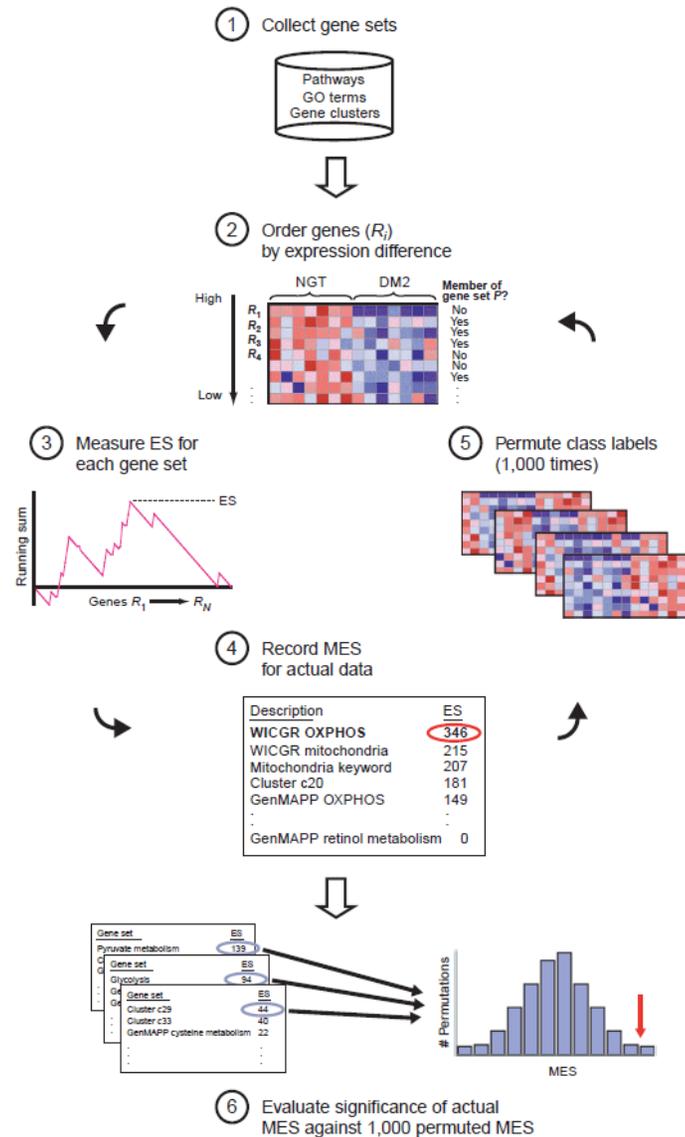
25 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_MF_2	oxidoreductase activity	RT	482	2.1	3.4E-314	2.2E-312	
<input type="checkbox"/>	GOTERM_MF_2	transferase activity	RT	597	2.6	5.2E-170	1.7E-168	
<input type="checkbox"/>	GOTERM_MF_2	cofactor binding	RT	195	0.9	2.4E-136	5.3E-135	
<input type="checkbox"/>	GOTERM_MF_2	lyase activity	RT	113	0.5	3.1E-74	5.1E-73	
<input type="checkbox"/>	GOTERM_MF_2	tetrapyrrole binding	RT	90	0.4	1.2E-54	1.5E-53	
<input type="checkbox"/>	GOTERM_MF_2	vitamin binding	RT	79	0.3	2.6E-41	2.9E-40	
<input type="checkbox"/>	GOTERM_MF_2	hydrolase activity	RT	456	2.0	3.0E-41	2.8E-40	
<input type="checkbox"/>	GOTERM_MF_2	carboxylic acid binding	RT	72	0.3	2.9E-30	2.3E-29	
<input type="checkbox"/>	GOTERM_MF_2	isomerase activity	RT	52	0.2	3.6E-16	2.4E-15	
<input type="checkbox"/>	GOTERM_MF_2	nucleoside binding	RT	284	1.2	4.3E-16	2.9E-15	
<input type="checkbox"/>	GOTERM_MF_2	oxygen binding	RT	27	0.1	1.0E-14	6.0E-14	
<input type="checkbox"/>	GOTERM_MF_2	peroxidase activity	RT	23	0.1	2.3E-14	1.2E-13	

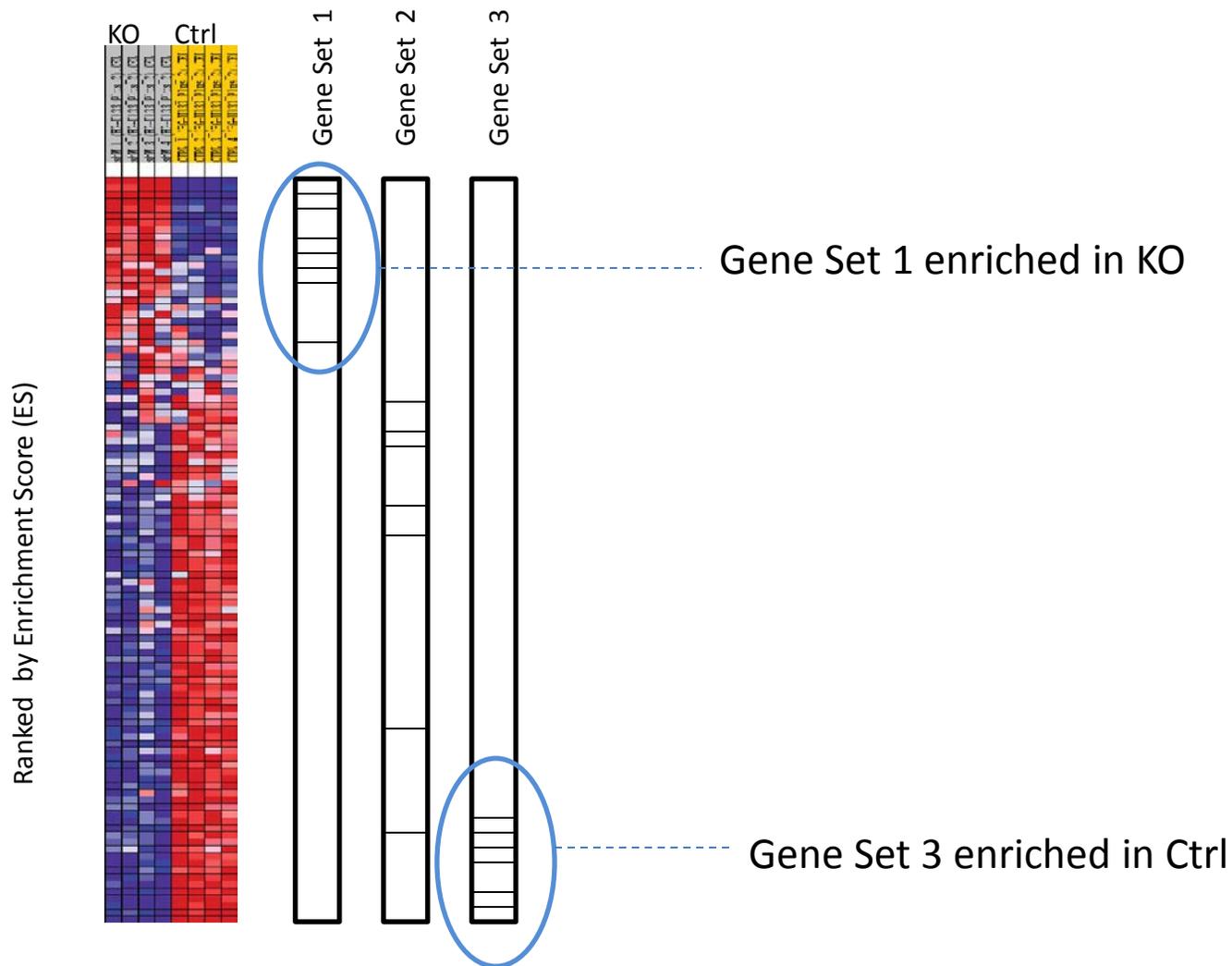
Expand "Options" to change parameters: eg. p-value correction by FDR or Benjamini

Select Database: eg. GO, Panther, etc. to expand

# Gene Set Enrichment Analysis (GSEA)



# GSEA





# GSEA: Interpreting Results

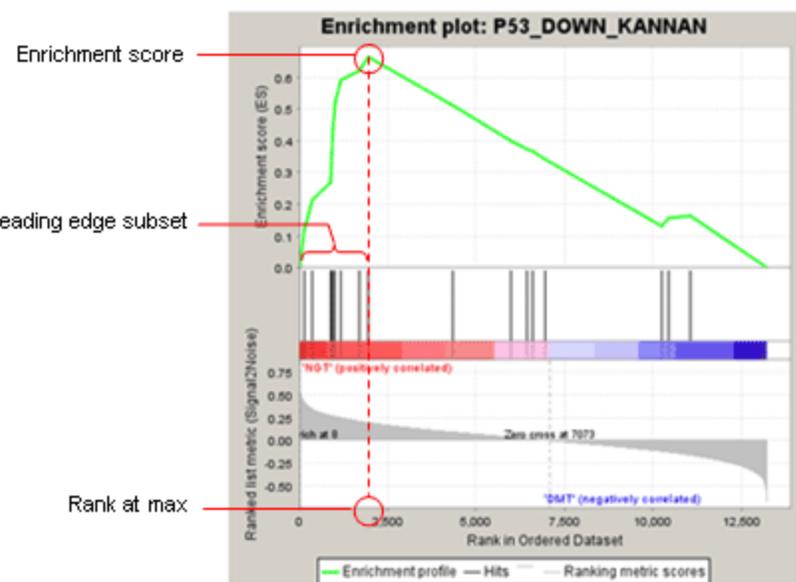
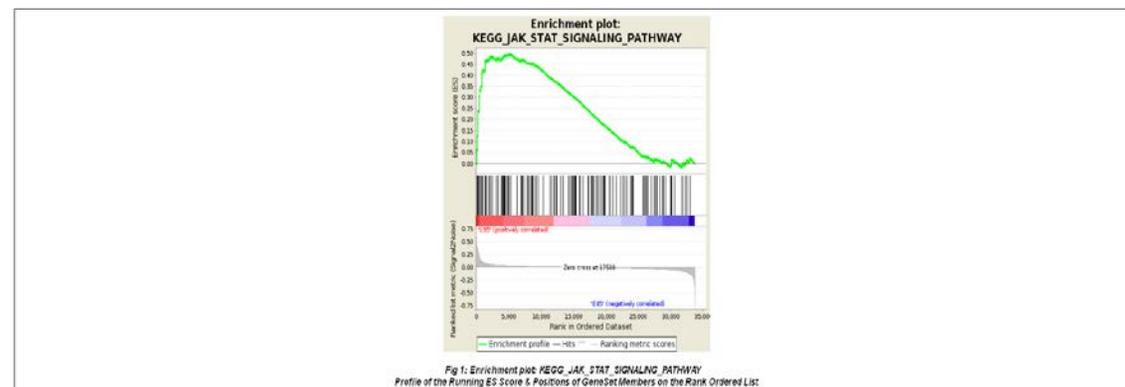
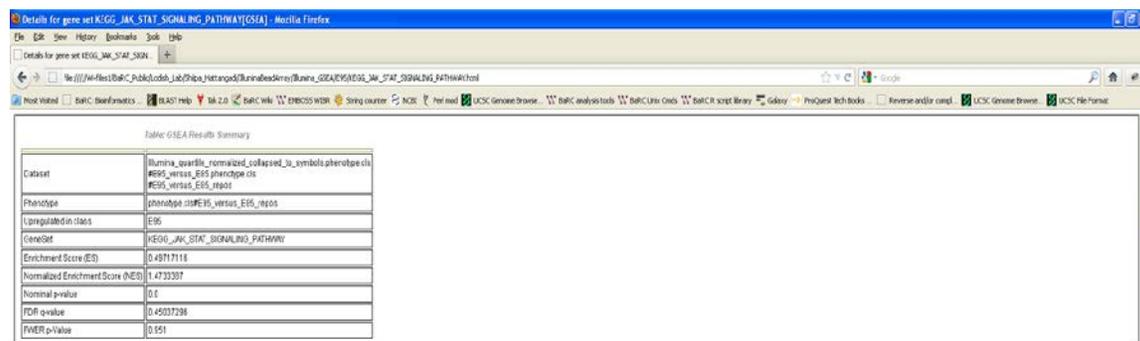


Table: GSEA Results (Data not shown)

RANK	GENE SYMBOL	GENE TITLE	RANK IN GENE LIST	RANK METRIC SCORE	RUNNING ES	ES POS. ENRICHMENT
1	CD81	CD81	140	0.391	0.0623	Yes
2	LXN2B	LXN2B	173	0.382	0.1250	No
3	OSM	OSM	256	0.348	0.1810	Yes
4	SOC1	SOC1	268	0.338	0.2385	No

# BiNGO

Cytoscape Desktop (New Session)

File Edit View Select Layout Plugins Help

Control Panel

Network VizMapper™ Editor Filters

Network Nodes Edges

Data Panel

BiNGO Settings

BiNGO settings

Save settings as default Help

Cluster name:

Get Cluster from Network  Paste Genes from Text

Do you want to assess over- or underrepresentation:

Overrepresentation  Underrepresentation

Visualization  No Visualization

Select a statistical test:

Hypergeometric test

Select a multiple testing correction:

Benjamini & Hochberg False Discovery Rate (FDR) correction

Choose a significance level:

0.05

Select the categories to be visualized:

Overrepresented categories after correction

Select reference set:

Use whole annotation as reference set

Select ontology file:

GO\_Molecular\_Function

Select namespace:

biological\_process

Select organism/annotation:

Homo Sapiens

Discard the following evidence codes:

Check box for saving Data Save BiNGO Data file in : Start BiNGO

Test: Hypergeometric or Binomial

p-value correction

Ref. or background, Including custom list

Gene Ontology

Welcome to Cytoscape 2.7.0 Right-click + drag to ZOOM Middle-click + drag to PAN



# BiNGO: Output

The screenshot displays the Cytoscape Desktop interface with a BiNGO output window. The network diagram shows a central 'binding' node connected to various other nodes representing biological processes. The nodes are color-coded based on their p-values, with a color scale ranging from yellow (5.00E-2) to red (<5.00E-7).

**BiNGO output window:**

GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
3824	catalytic activity	1.2263E-79	6.1562E-77	170/172 98.0%	5088/15430 33.0%	CYP24A1 GDA PTGS2 GNPD2 SQMS2 ADCY7 PTGS1 DSE TPK1 GLT8D1 GLT8D2 PTGIS CH25H ELOVL2 PDE4B ST3GAL6 GXYLT2 NMNAT2 SPTLC...
16491	oxidoreductase activity	2.0371E-29	5.1132E-27	52/172 30.2%	686/15430 4.5%	ACOX2 CYBR3 CYP24A1 CYBR2 PTGS2 PTGS1 AKR1C2 FAR2 TDO2 MSRA PTGIS HMOX1 CH25H GPX3 GPX7 AKR1C1 MICAL2 CBR3 CD...
16740	transferase activity	1.7448E-20	2.9196E-18	64/172 37.2%	1635/15430 10.6%	SGMS2 GLT8D1 TPK1 ST6GALNACS GLT8D2 GLT2SD2 ELOVL2 ST3GAL6 AGPAT9 DSEL GXYLT2 AGPAT4 NMNAT2 SPTLC3 HSDC1 UGCG CH5T2 P...
16757	transferase activity, transferring glycosyl groups	2.2417E-19	2.8133E-17	27/172 15.6%	246/15430 1.6%	GALNT2 GCNT3 GALNT5 ST6SIA1 UPP1 B3GNT9 GLT8D1 ST6GALNACS GLT8D2 UGT1A6 GLT2SD2 GALNT10 ST3GAL6 GXYLT2 GALNT14 B3GALT...
8194	UDP-glycosyltransferase activity	1.1983E-16	1.2031E-14	19/172 11.0%	124/15430 0.8%	GALNT2 GCNT3 B3GALT2 GALNT5 UGCG GALNT11 CSGLCA-T CS6GALNACT1 UGT1A6 GALNT10 GLT2SD2 CH5Y3 GXYLT2 HAS2 CH5Y1 HAS3 EXT1...
16758	transferase activity, transferring hexosyl groups	3.9867E-15	3.3355E-13	20/172 11.6%	171/15430 1.1%	GALNT2 GCNT3 B3GALT2 GALNT5 UGCG GALNT11 CSGLCA-T CS6GALNACT1 B3GNT9 UGT1A6 GALNT10 GRE1 GLT2SD2 CH5Y3 HAS2 CH5Y1 HAS...
5506	iron ion binding	1.5320E-11	1.0986E-9	18/172 10.4%	206/15430 1.3%	NOX4 CYP2U1 CYP24A1 CYP1B1 PTGS2 PTGS1 FAD53 CYP4F11 CD01 CPS1 CYP27C1 TDO2 PTGIS CHEM HMOX1 AOX1 GUCY1B3 NTSE...
16788	hydrolase activity, acting on ester bonds	3.6619E-11	2.2979E-9	31/172 18.0%	699/15430 4.5%	LRP4 ARSE ENPP1 ENPP2 ARS3 ARSK ACOT9 PLC2 PLCB4 RCE RDE1C HMOX1 GALNS PDE4B SYN3 PPAR2A PPAR2B NTSE NCEH1 B5T1 PPA...
20037	heme binding	8.0840E-11	4.5091E-9	14/172 8.1%	122/15430 0.8%	CYP2U1 NOX4 CYP24A1 CYP1B1 PTGS2 PTGS1 FAD53 CYP4F11 CPS1 CYP27C1 TDO2 PTGIS HMOX1 GUCY1B3
46906	tetrapyrrole binding	1.9135E-10	9.6058E-9	14/172 8.1%	130/15430 0.9%	CYP2U1 NOX4 CYP24A1 CYP1B1 PTGS2 PTGS1 FAD53 CYP4F11 CPS1 CYP27C1 TDO2 PTGIS HMOX1 GUCY1B3



# GeneGO

Most Popular Questions | Upload | Workflows & Reports | **One-click Analysis** | Build Network | Custom Content | Predict Compound Activity (MetaDrug) | Search & Browse Content

**Enrichment Ontologies** [?](#)  
Scores and ranks entities in functional ontologies most relevant in activated dataset(s).

**GeneGO Ontologies**

- GeneGO Pathway Maps**
- GeneGO Map Folders
- GeneGO Process Networks
- GeneGO Diseases (by Biomarkers)
- GeneGO Disease Biomarker Networks
- GeneGO Drug Target Networks
- GeneGO Toxic Pathologies
- GeneGO Drug and Xenobiotic Metabolism Enzymes
- GeneGO Toxicity Networks
- GeneGO Metabolic Networks
- GeneGO Metabolic Networks (Endogenous)

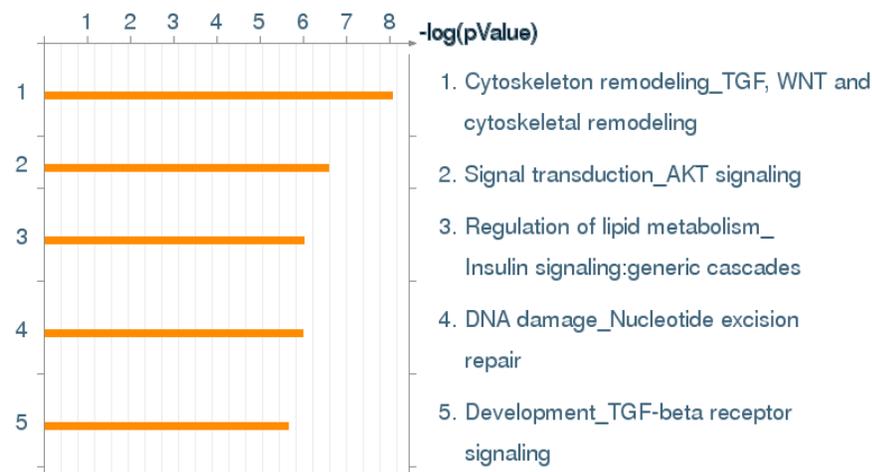
**Public Ontologies**

- GO Processes
- GO Molecular Functions

**Interactome** [?](#)  
Detailed analysis of interaction space for activated datasets and gene lists

- Interactions by Protein Function
- Transcription Factors
- Significant Interactions Within Set(s)
- Interactome Topology
- Enrichment by Protein Function
- Interactions Between Datasets (all)
- Interactions Between Datasets (TR)
- Drug Lookup for Your Gene Lists and Datasets [?](#)

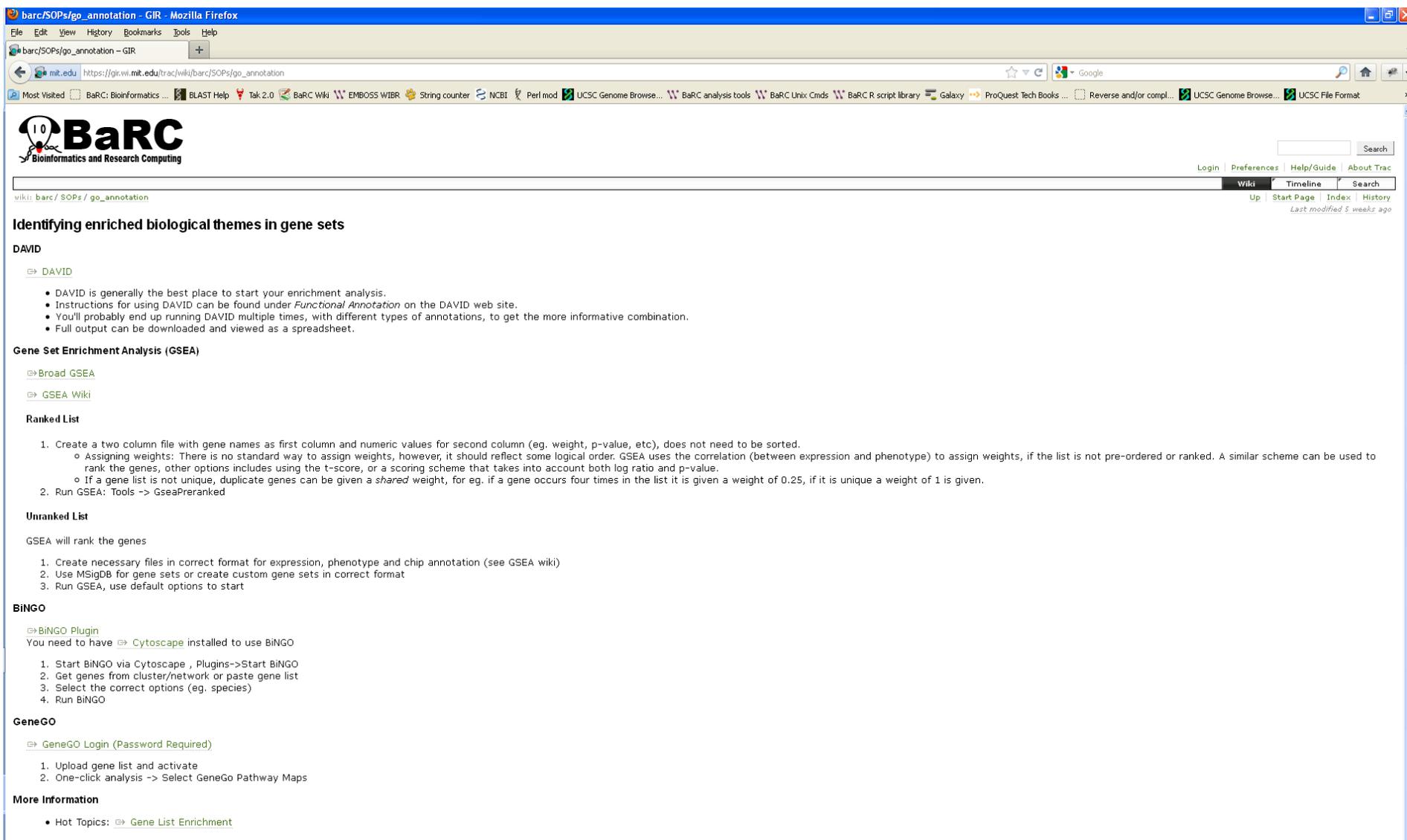
[Hide Description](#)





# BaRC SOP

[https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go\\_annotation](https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go_annotation)



The screenshot shows a Mozilla Firefox browser window with the address bar displaying [https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go\\_annotation](https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go_annotation). The page content includes the BaRC logo, navigation links (Login, Preferences, Help/Guide, About Trac), and a search bar. The main content is titled "Identifying enriched biological themes in gene sets" and is categorized under "DAVID".

**Identifying enriched biological themes in gene sets**

**DAVID**

⇒ DAVID

- DAVID is generally the best place to start your enrichment analysis.
- Instructions for using DAVID can be found under *Functional Annotation* on the DAVID web site.
- You'll probably end up running DAVID multiple times, with different types of annotations, to get the more informative combination.
- Full output can be downloaded and viewed as a spreadsheet.

**Gene Set Enrichment Analysis (GSEA)**

⇒ Broad GSEA

⇒ GSEA Wiki

**Ranked List**

1. Create a two column file with gene names as first column and numeric values for second column (eg. weight, p-value, etc), does not need to be sorted.
  - Assigning weights: There is no standard way to assign weights, however, it should reflect some logical order. GSEA uses the correlation (between expression and phenotype) to assign weights, if the list is not pre-ordered or ranked. A similar scheme can be used to rank the genes, other options includes using the t-score, or a scoring scheme that takes into account both log ratio and p-value.
  - If a gene list is not unique, duplicate genes can be given a *shared* weight, for eg. if a gene occurs four times in the list it is given a weight of 0.25, if it is unique a weight of 1 is given.
2. Run GSEA: Tools -> GseaPreranked

**Unranked List**

GSEA will rank the genes

1. Create necessary files in correct format for expression, phenotype and chip annotation (see GSEA wiki)
2. Use MSigDB for gene sets or create custom gene sets in correct format
3. Run GSEA, use default options to start

**BINGO**

⇒ BINGO Plugin

You need to have ⇒ Cytoscape installed to use BINGO

1. Start BINGO via Cytoscape , Plugins->Start BINGO
2. Get genes from cluster/network or paste gene list
3. Select the correct options (eg. species)
4. Run BINGO

**GeneGO**

⇒ GeneGO Login (Password Required)

1. Upload gene list and activate
2. One-click analysis -> Select GeneGo Pathway Maps

**More Information**

- Hot Topics: ⇒ Gene List Enrichment

# Enrichment Analysis: Practicalities

- Choose a tool that
  - includes your species
  - includes your genes or identifiers
  - has up-to-date annotation
  - allows user-defined background
- Try a few tools
- Use gene lists with varying lengths (ie. different thresholds)
- Ignore enriched categories which,
  - contain very few genes
  - highly overlap with other categories
- Graphical or text summary

## Further Reading

- Clark, N.R., and Ma'ayan, A. *Introduction to Statistical Methods for Analyzing Large Data Sets: Gene-Set Enrichment Analysis* Science Signaling 4:190 (2011)
- Huang, D.W., et al. *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists* NAR 37:1-13 (2008)
- Rivals, et. al *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics 23:401-407 (2006)

# Supplementary: Fisher's Exact Test in R

Are the genes in my list from adipose tissue enriched for fatty acid (FA) cycle?

	Number of Genes in Gene Set	Number of Genes not in Gene Set	Total
Genes in (my) Gene List	$a$	$c$	$a+c$
All Genes	$b$	$d$	$b+d$
Total	$a+b$	$c+d$	$a+b+c+d$

2x2 contingency table

	All Tissues	Adipose
Amino Acid	35	
Bile Acid	6	
Carbohydrate Storage	13	2
Cholesterol	14	
CoA	14	
Cofactor	24	
Creatine	5	
Cysteine	9	1
Detox	28	2
<b>Fatty Acid</b>	<b>120</b>	<b>7</b>
Folate	11	
Glutamate	10	
Glycan Degradation	46	
Glycan Sulfate	20	1

:

Urea	6	
Vitamin A	16	1
Vitamin B6	4	
Total	1663	27

# Supplementary: Fisher's Exact Test in R

	FA cycle	Metabolic
Adipose	7	27
All Tissues	120	1663

	FA cycle	Non-FA pathway	(Row) Total
Adipose	7	20	27
Non-adipose	113	1523	1636
(Column) Total	120	1543	1663

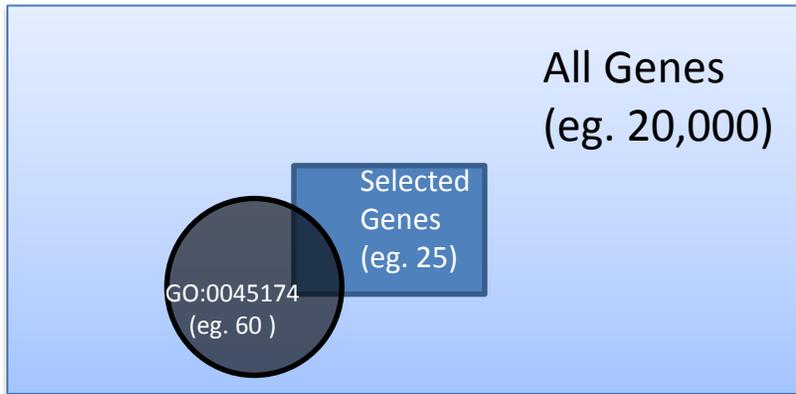
```
R session:
> myData <-matrix(c(7,113,20,1523),nr=2)
> fisher.test(myData, alternative="greater")
```

Fisher's Exact Test for Count Data

```
data: myData
p-value = 0.002253
```

# Supplementary: Hypergeometric Distribution

What is the probability of observing 10 selected/significant genes in the GO Term?



```
#R Commands
#null hypergeometric distribution
>dhyper(0:25,60,19940,25) #see fig. on right
>sum(dhyper(10:25,60,19940,25))
#p-value: 8.459171e-20

#Alternative using Fisher's Exact Test
#           In GO   Not in GO
#       Sig.    10    15
#   Non-Sig.   50   19925
#
>myContingencyTable<-matrix(c(10,50,15,19925),nr=2)
>fisher.test(myContingencyTable)
#p-value: 8.459171e-20
```

Hypergeometric Null Distribution

