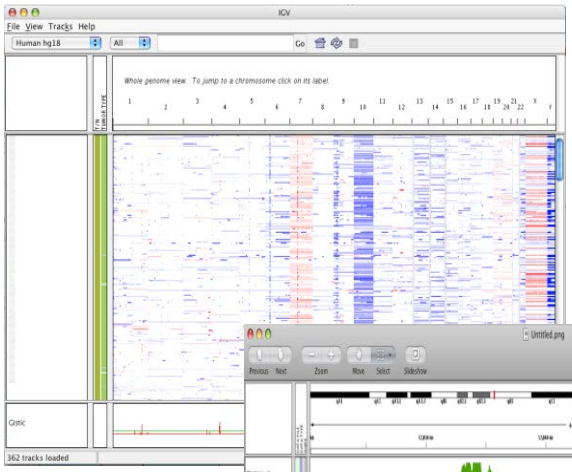
Enrichment Analysis

Prat Thiru

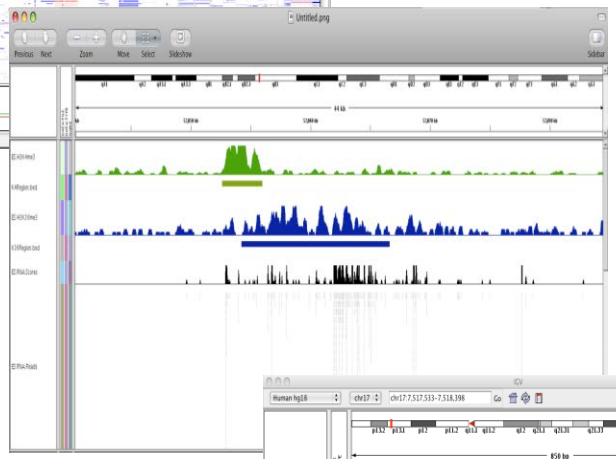
Outline

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

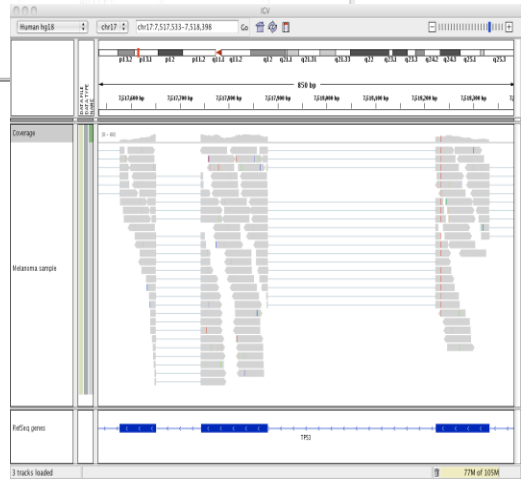
Overview



Microarrays



Epigenomics



RNA-Seq

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

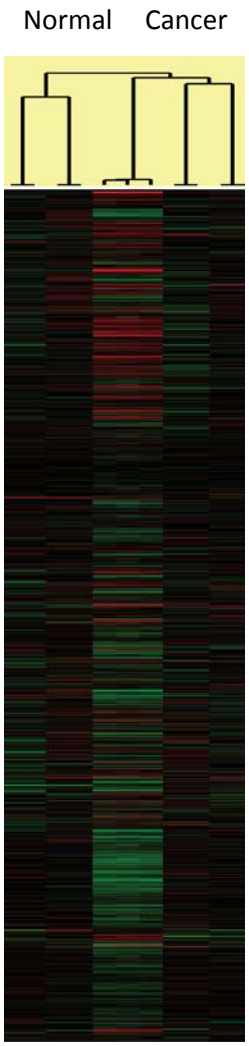
Long List of Genes

?

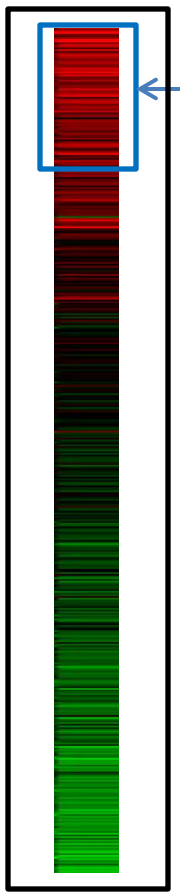
Goals of Enrichment Analysis

- 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



Sort/Rank



Use a subset of genes and see if these are functionally enriched

- Need to choose threshold or cutoff

Use all genes and see if these are functionally enriched

- No need to choose a cutoff
- Gene Set Enrichment Analysis (GSEA)

Databases for Annotations

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

*a collection of databases

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

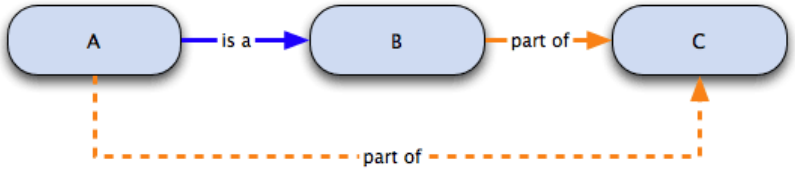
Gene Ontology (GO)

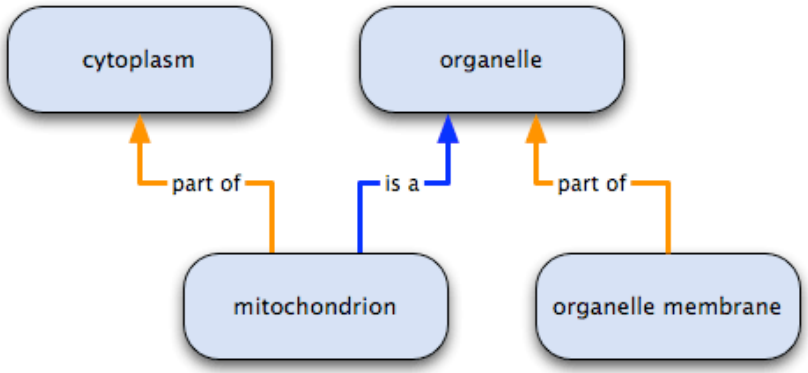
- Collection of gene sets with controlled vocabulary
- Cellular Component: parts of a cell (eg. nucleus, ER)
- Molecular Function: activity of a gene product (eg. binding, catalysis)
- Biological Process: series of events accomplished by one or more ordered assemblies of molecular functions (eg. pyrimidine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

http://www.geneontology.org/GO.doc.shtml

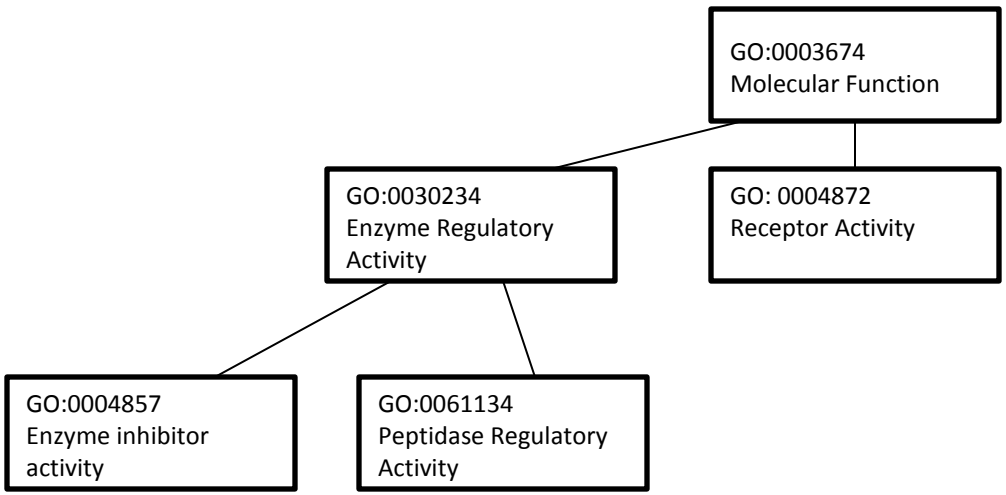
Gene Ontology

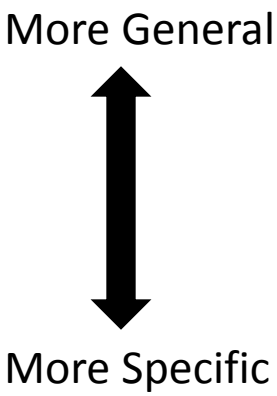
- Based on “is a” or “part of” relationship



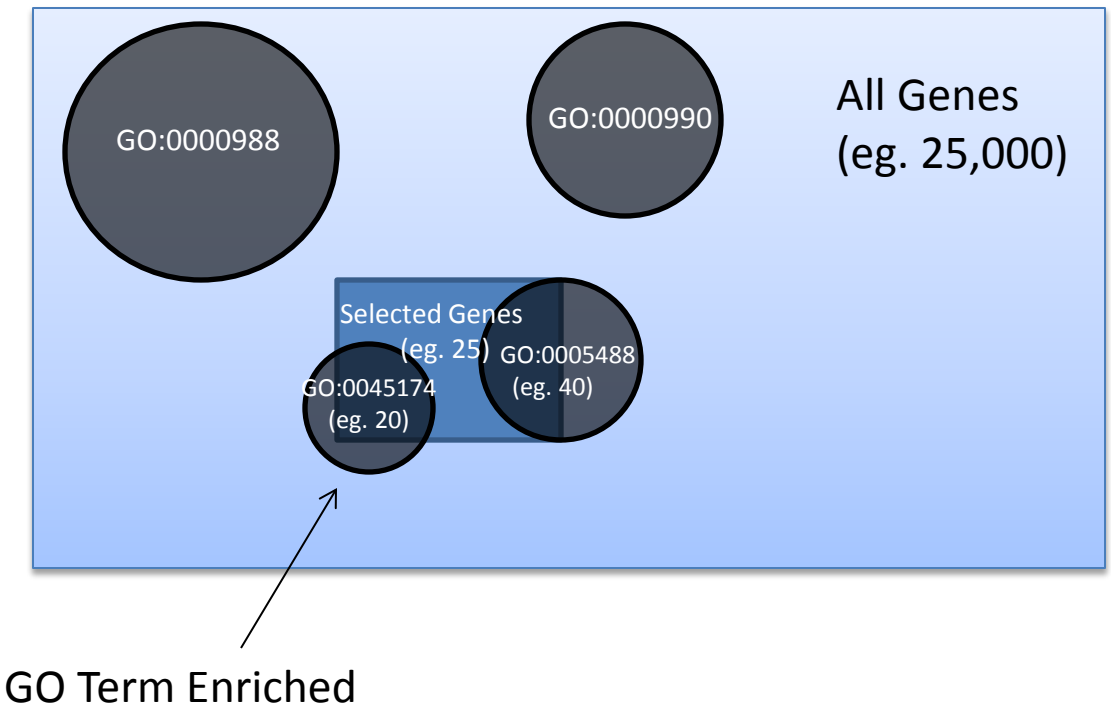


- Hierarchical relationship



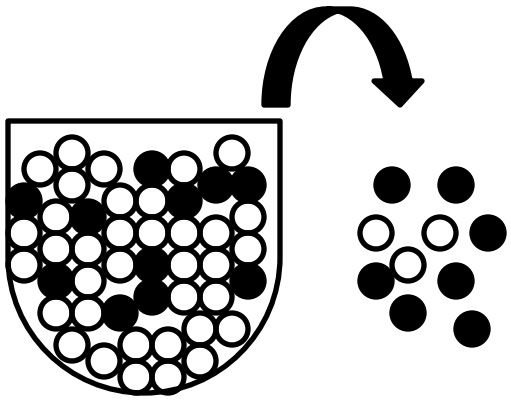


GO Enrichment



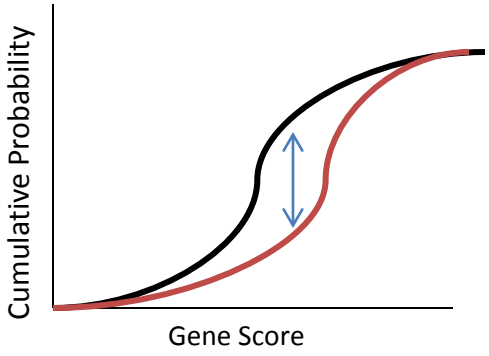
Assessing Significance of Enrichment

- Fisher's Exact Test (Hypergeometric Test)



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

- Kolmogorov-Smirnov (KS) Test

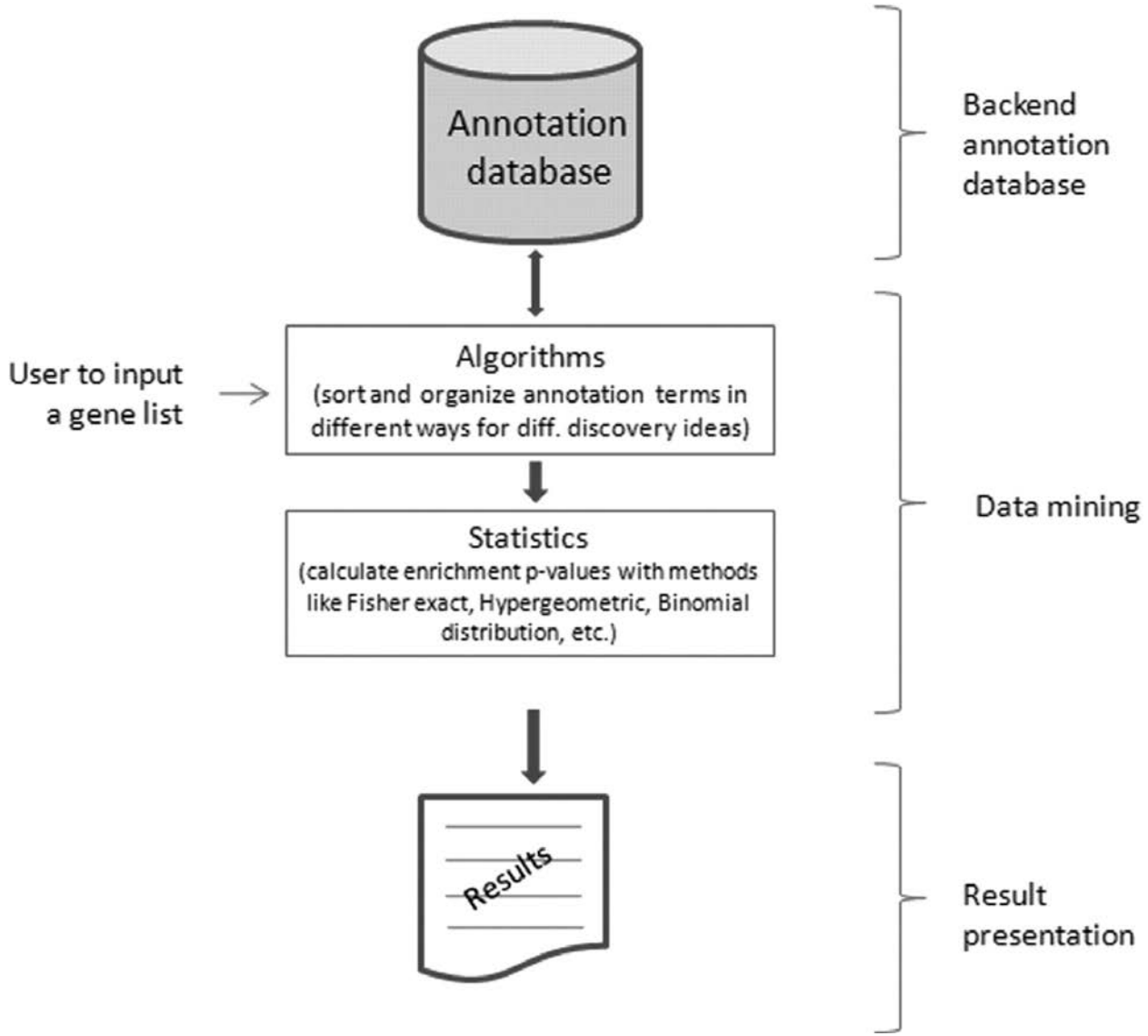


Assessing Significance of Enrichment: Which test?

- 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

Rivals, et. al *Enrichment or depletion of a GO category within a class of genes: which test?* Bioinformatics 23:401-407 (2006)

Enrichment Analysis Tool Infrastructure



Huang, D.W., et al. *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists* NAR 37:1-13 (2008)

Enrichment Analysis Tools

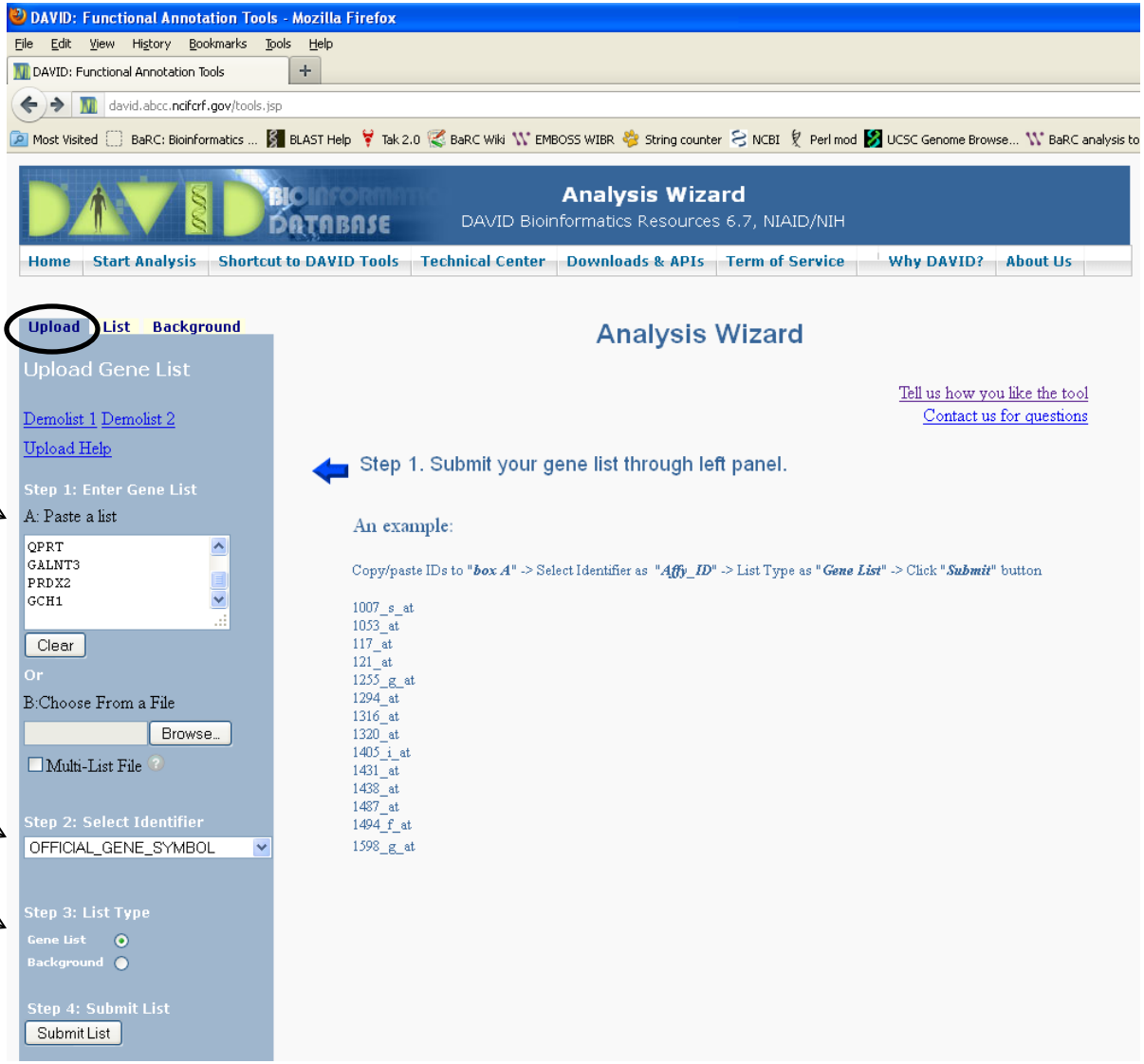
Tool	Statistical Method	Website
DAVID	Fisher	http://david.abcc.ncifcrf.gov
GSEA	KS Test	http://www.broadinstitute.org/gsea
BiNGO (Cytoscape Plugin)	Hypergeometric; Binomial	http://www.psb.ugent.be/cbd/papers/BiNGO/Home.html
GeneGO*	Hypergeometric	http://www.genego.com
GoMiner	Fisher	http://discover.nci.nih.gov/gominer/index.jsp

*commercial

Enrichment Analysis: Factors to Consider

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

DAVID

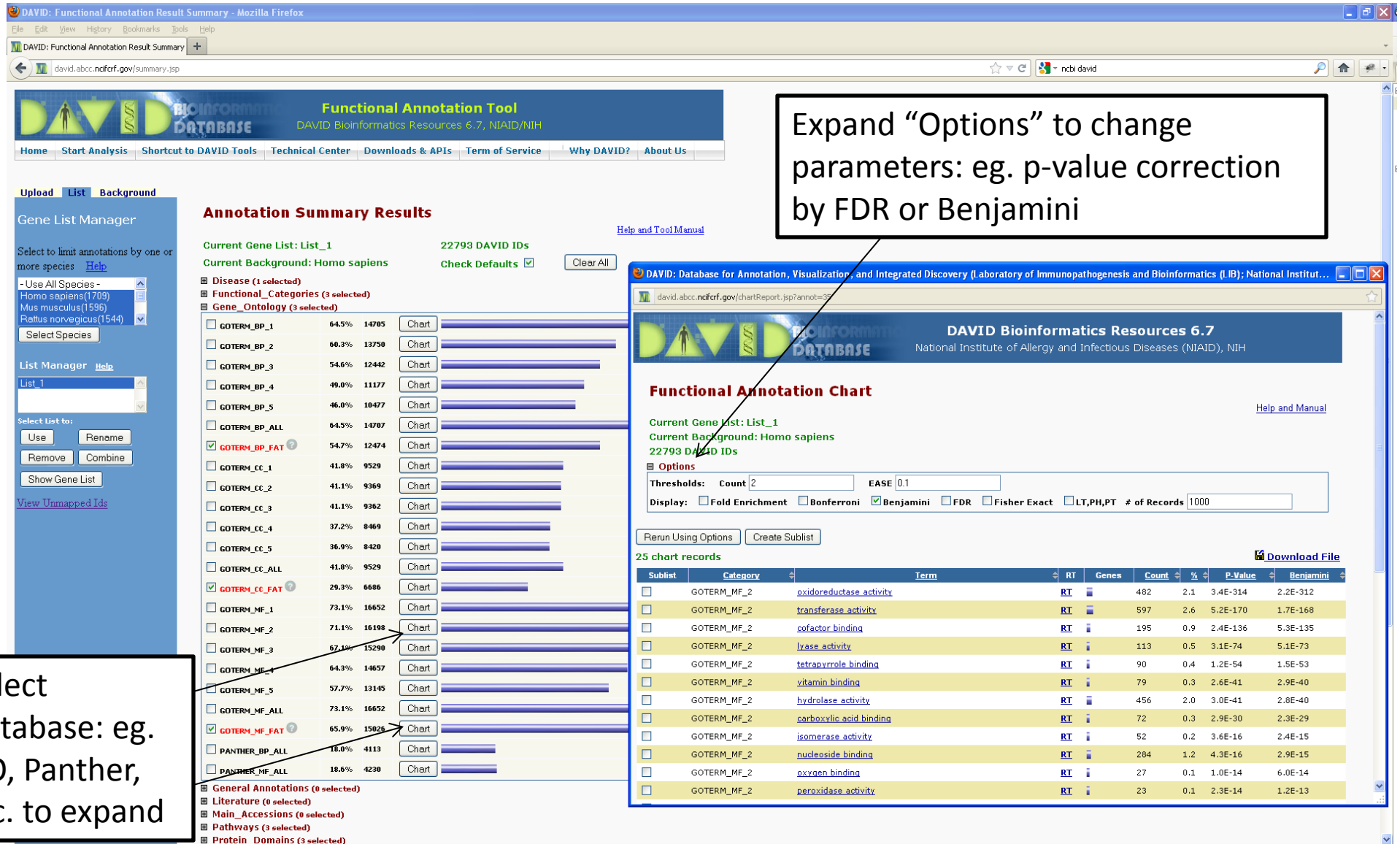


Gene List

Choose Identifier

Select List Type

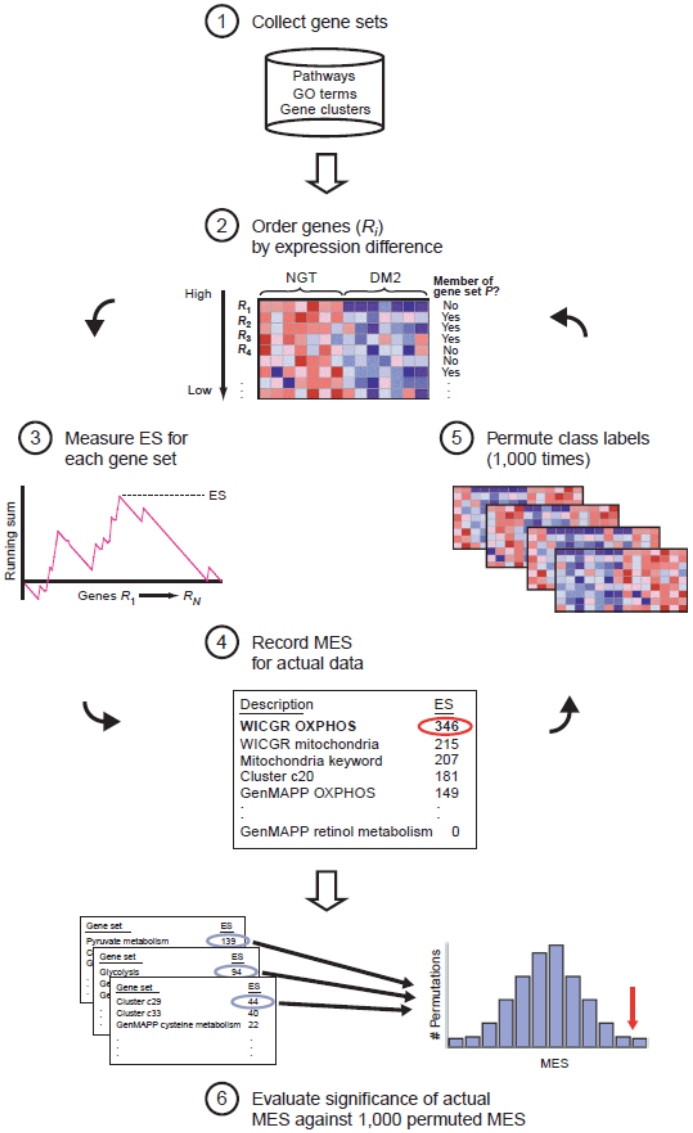
DAVID: Output



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

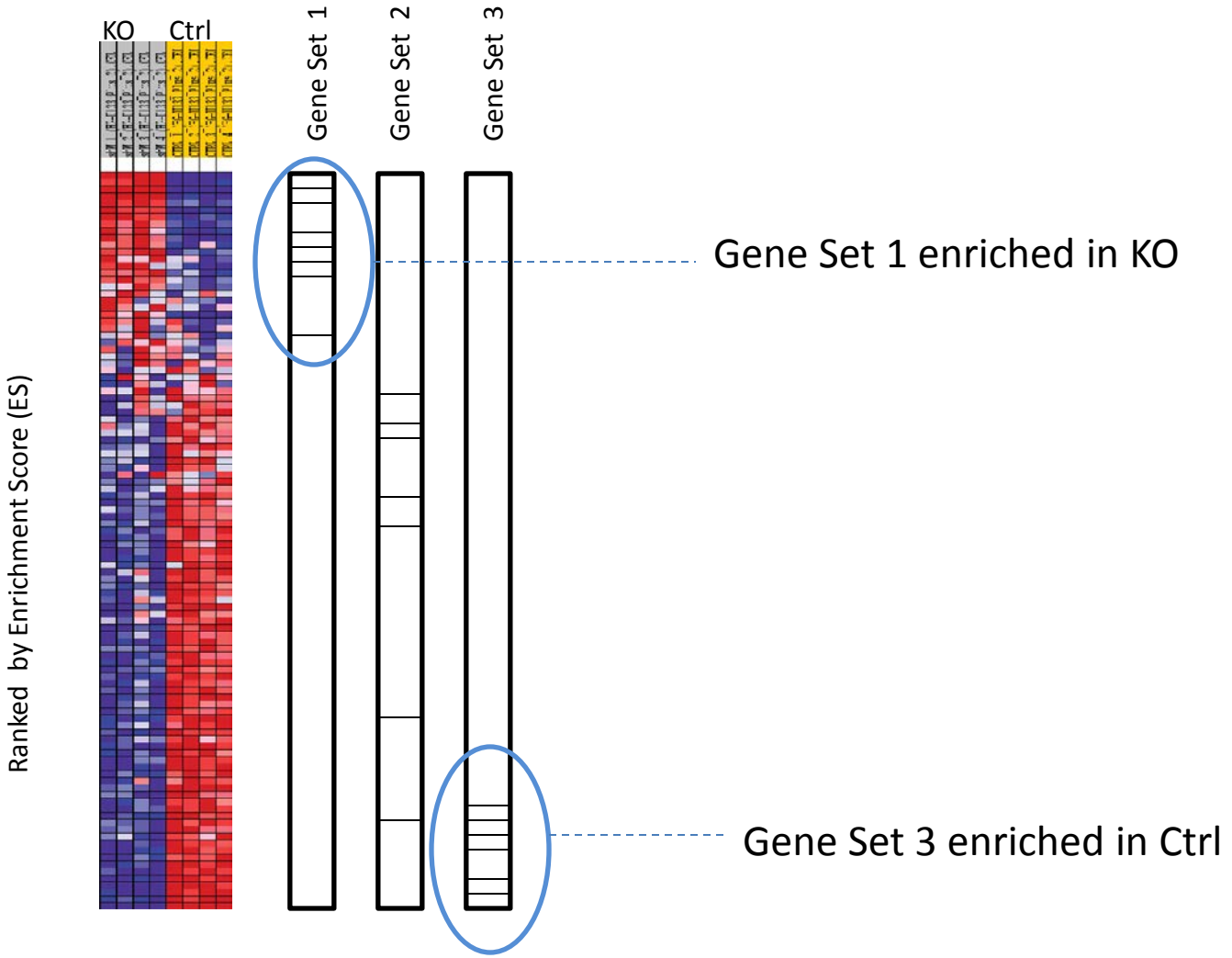
Select Database: eg. GO, Panther, etc. to expand

Gene Set Enrichment Analysis (GSEA)



Mootha, V.K., et al. *PGC-1α responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes* Nature Genetics 34(3):267-273 (2003)

GSEA


KO
Ctrl
Gene Set 1
Gene Set 2
Gene Set 3
Ranked by Enrichment Score (ES)
Gene Set 1 enriched in KO
Gene Set 3 enriched in Ctrl

GSEA: Interpreting Results

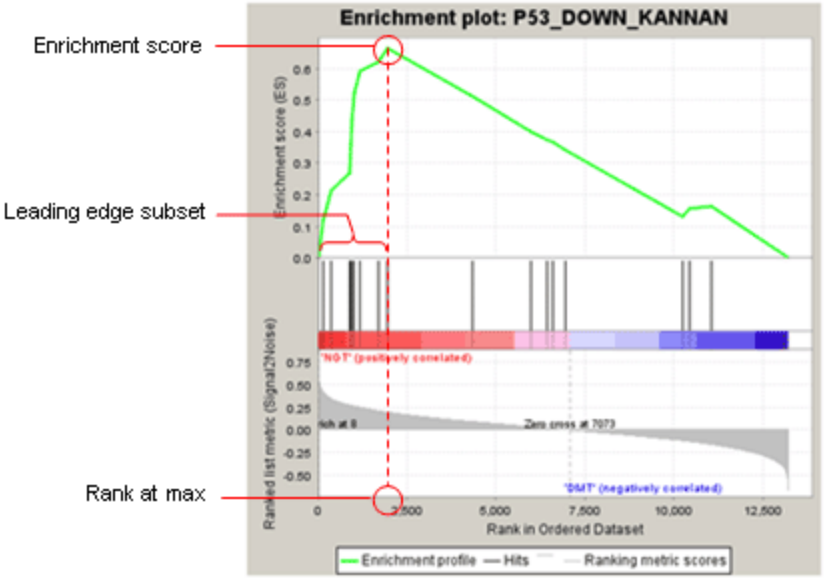
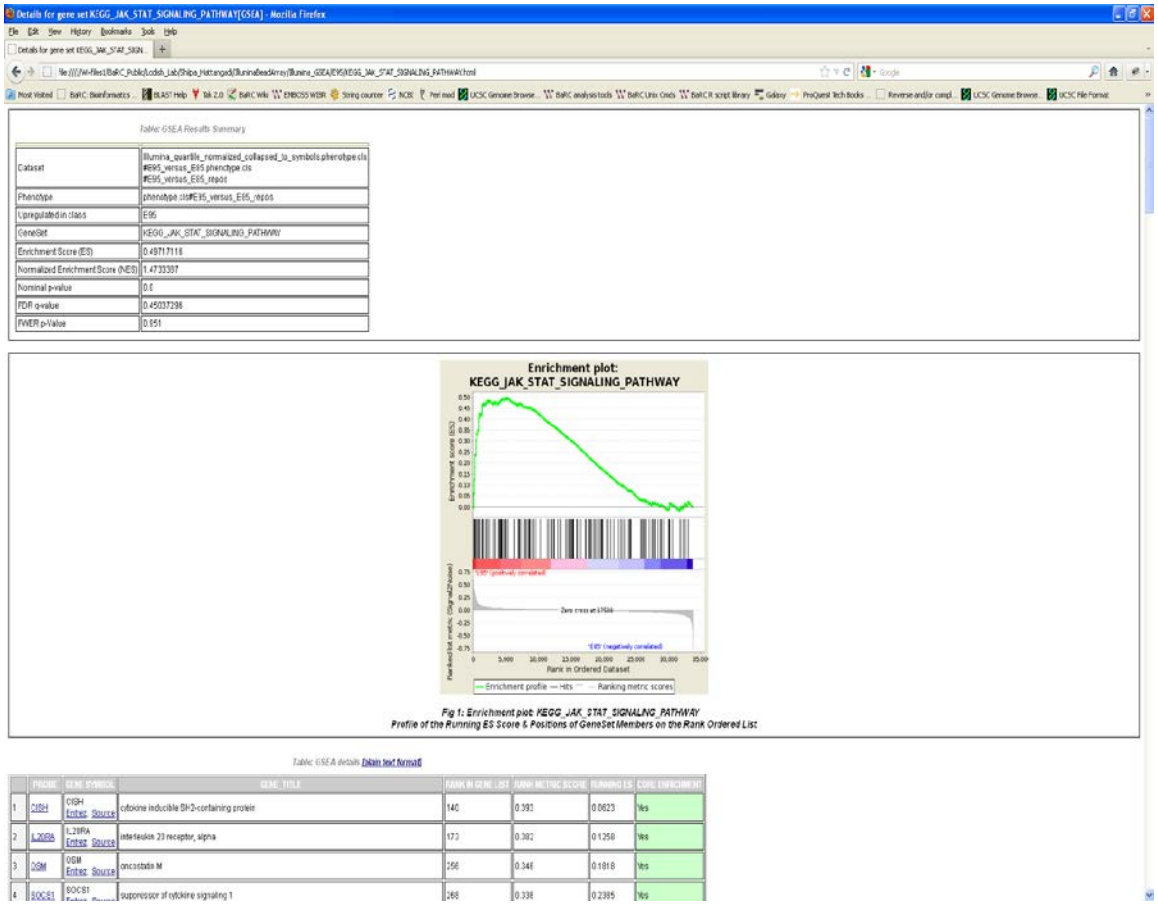
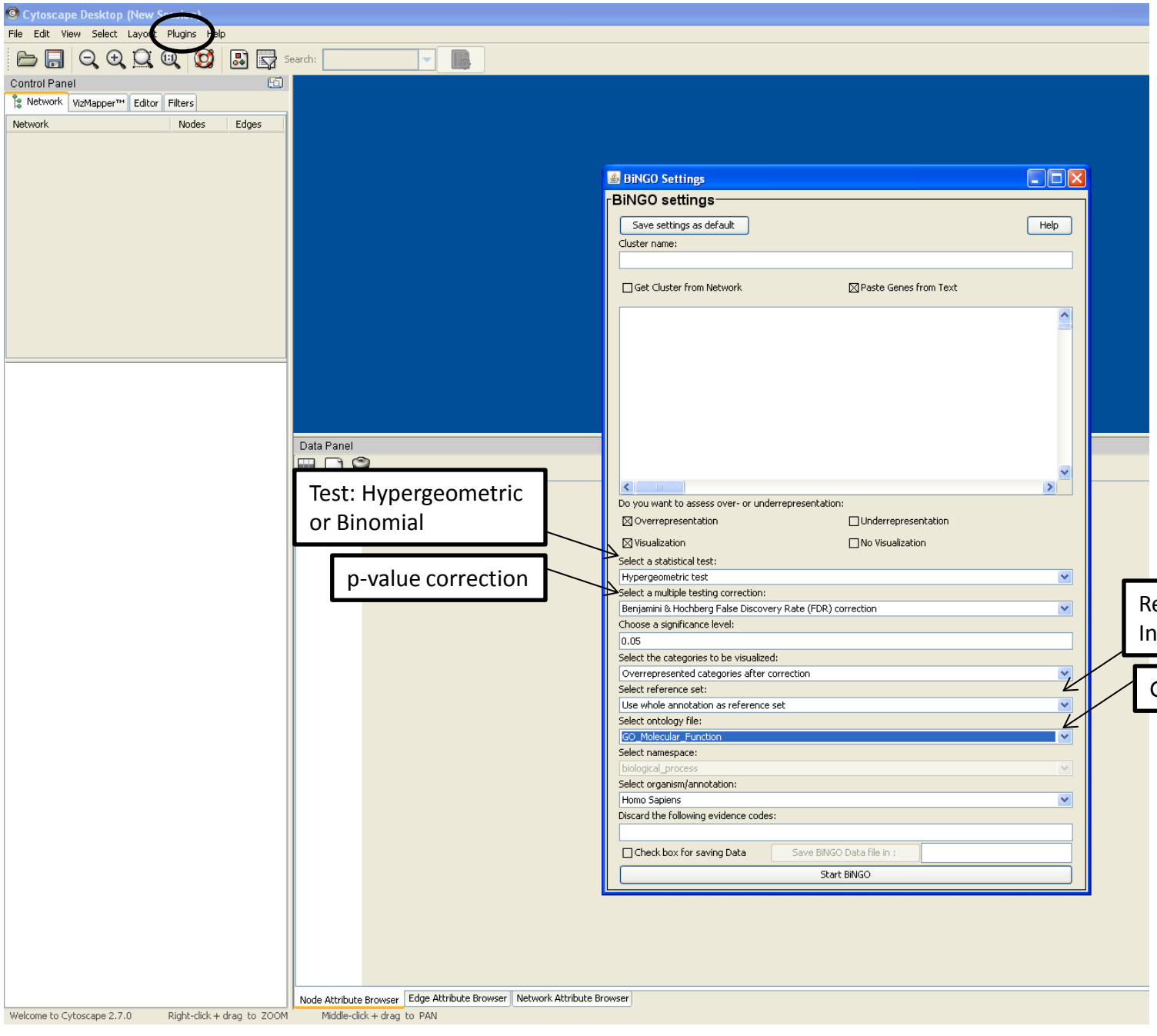


Fig 1: Enrichment plot: P53_DOWN_KANNAN
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

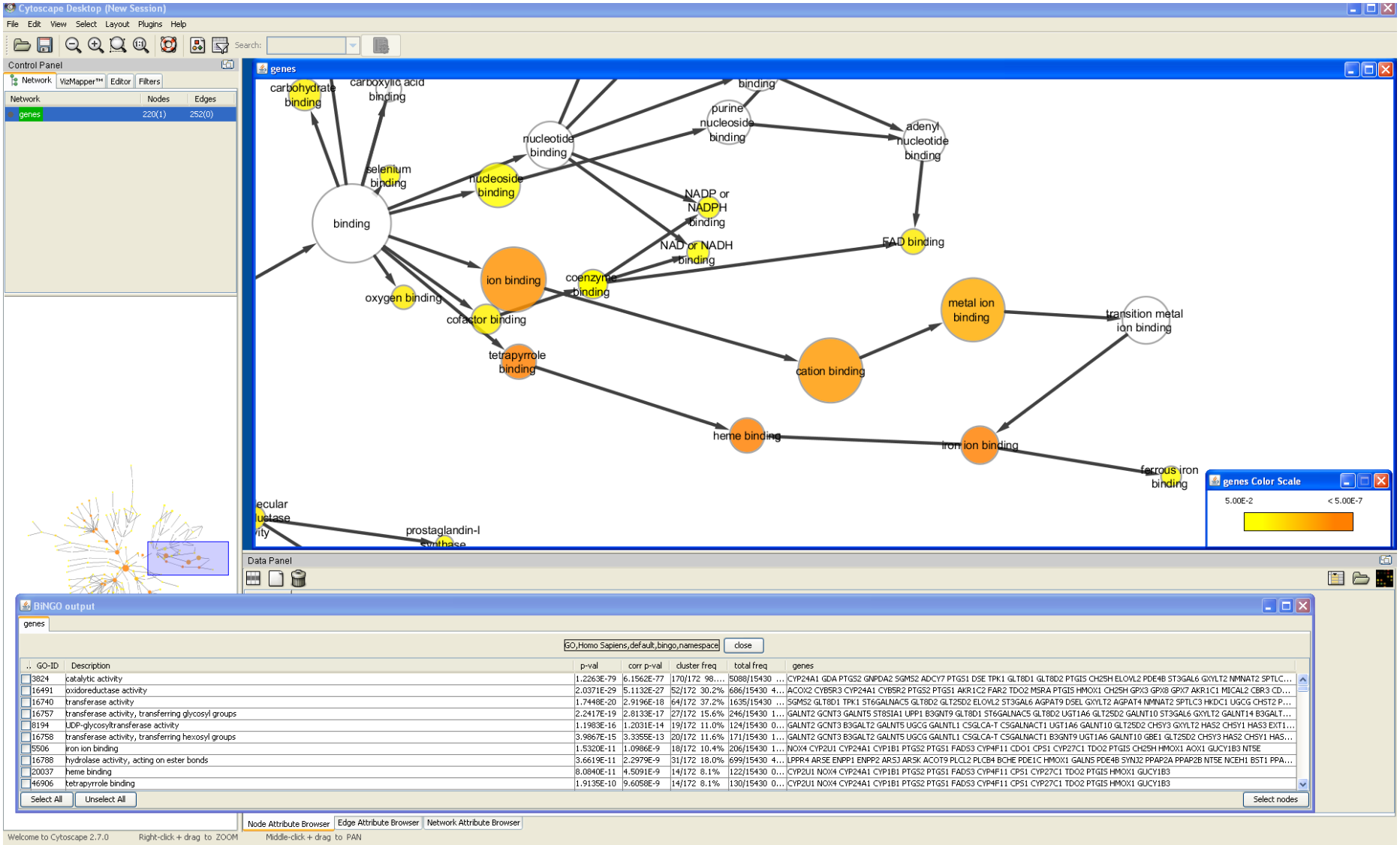


http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

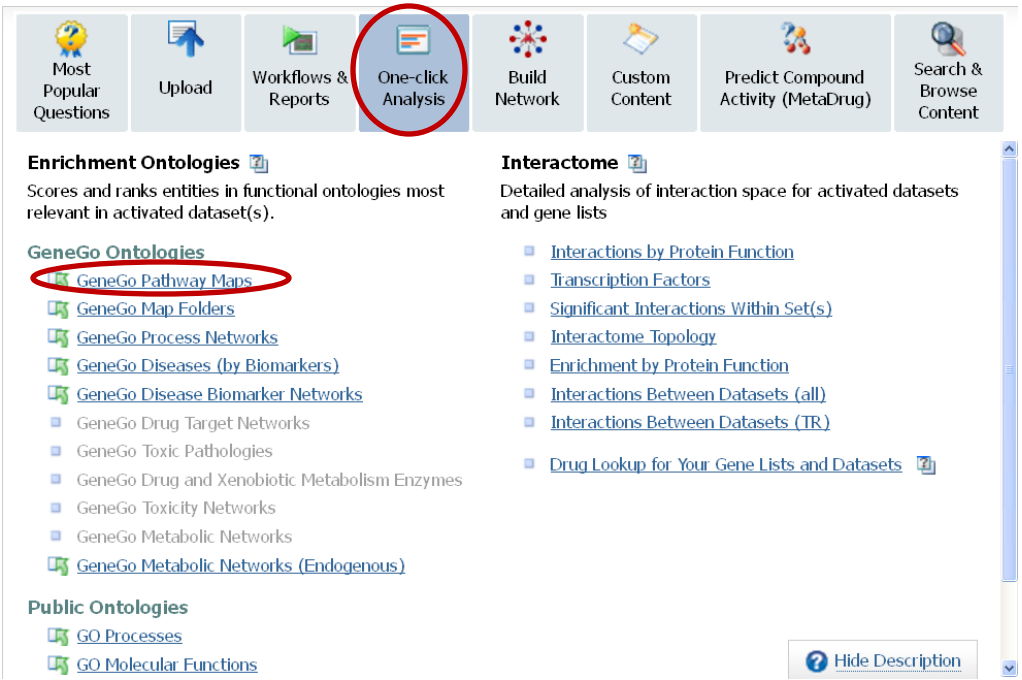
BiNGO

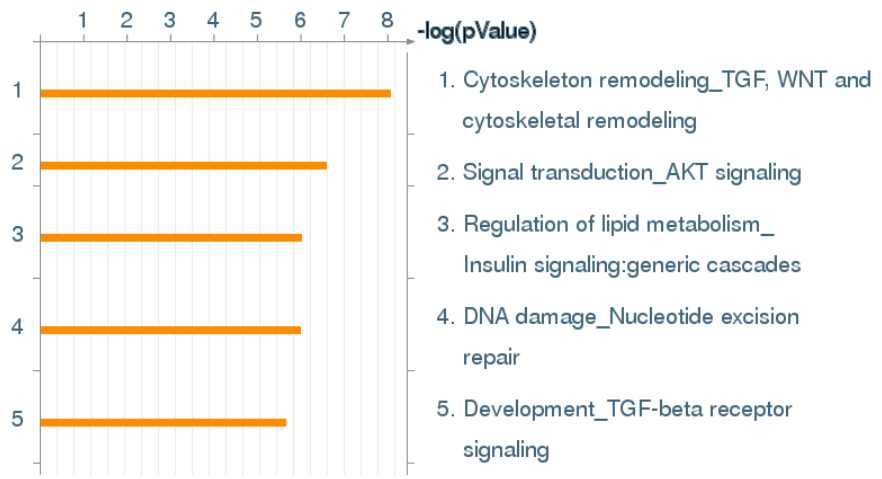

Cytoscape Desktop (New Session)
File Edit View Select Layout Plugins Help
Search:
Control Panel
Network VizMapper™ Editor Filters
Network Nodes Edges
Data Panel
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.7.0
Right-click + drag to ZOOM
Middle-click + drag to PAN
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

BiNGO: Output



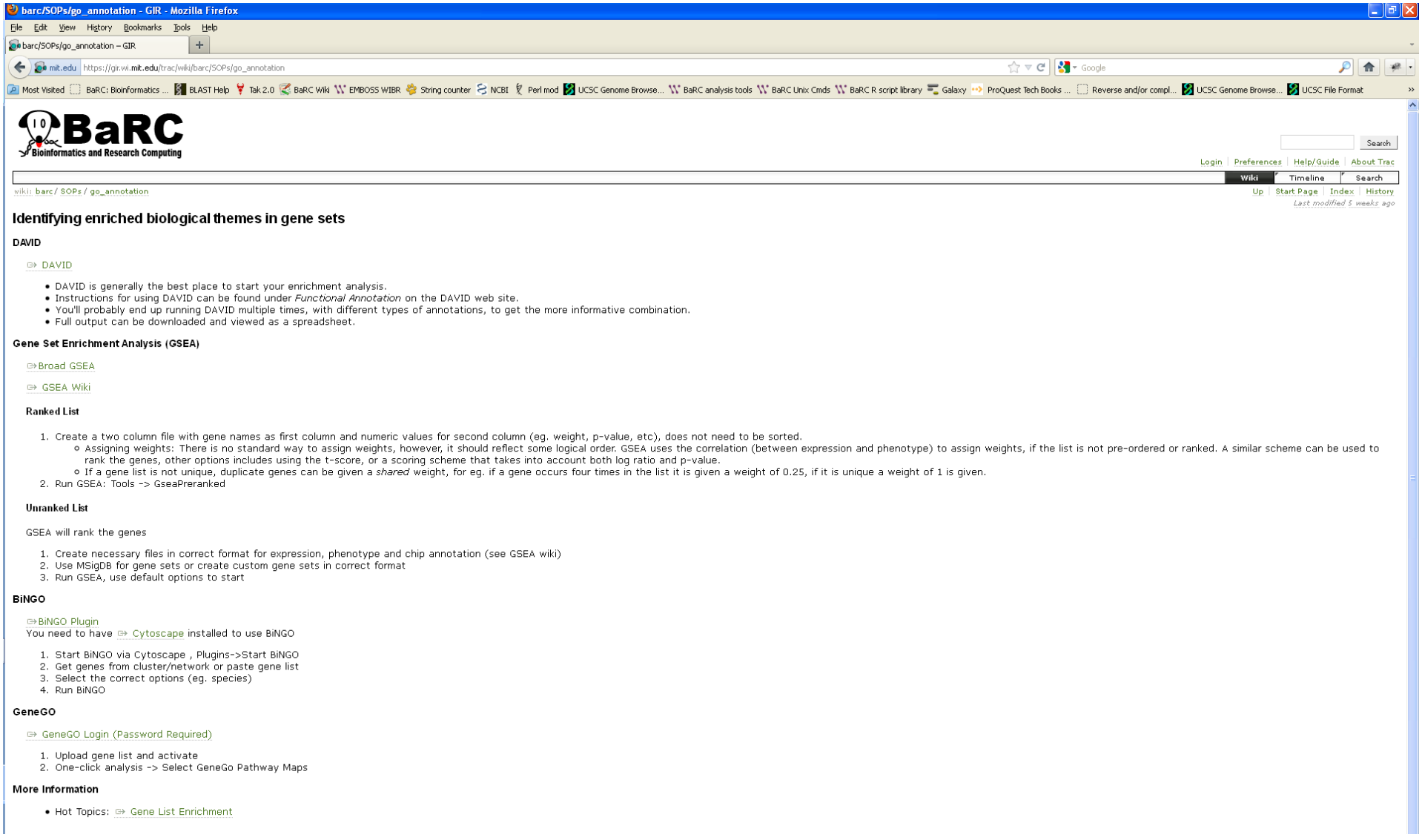
GeneGO


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


1 2 3 4 5 6 7 8
-log(pValue)
1
2
3
4
5
1. Cytoskeleton remodeling_TGF, WNT and cytoskeletal remodeling
2. Signal transduction_AKT signaling
3. Regulation of lipid metabolism_ Insulin signaling:generic cascades
4. DNA damage_Nucleotide excision repair
5. Development_TGF-beta receptor signaling

BaRC SOP

https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go_annotation



barc/SOPs/go_annotation - GIR - Mozilla Firefox

File Edit View History Bookmarks Tools Help

barc/SOPs/go_annotation – GIR

https://gir.wi.mit.edu/trac/wiki/barc/SOPs/go_annotation

BaRC Bioinformatics and Research Computing

Search

Login | Preferences | Help/Guide | About Trac

Wiki | Timeline | Search

wiki: barc / SOPs / go_annotation

Up | Start Page | Index | History

Last modified 5 weeks ago

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

http://jura.wi.mit.edu/bio/education/hot_topics/enrichment/Gene_list_enrichment_Mar10.pdf

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
Fatty Acid	**120**	**7**
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

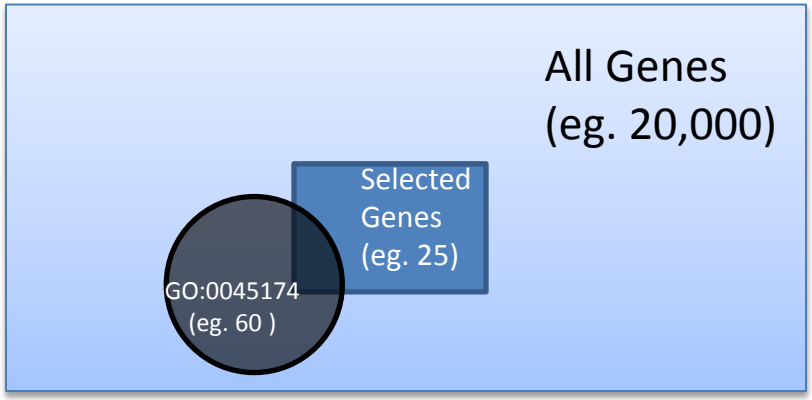
```
R session:
> myData <-matrix(c(7,113,20,1523),nr=2)
> fisher.test(myData, alternative="greater")

        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