



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



RNA-Seq

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	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/gse a/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.





Gene Ontology

• Based on "is a" or "part of" 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

ТооІ	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/Bi NGO/Home.html
GeneGO*	Hypergeometric	http://www.genego.com
GoMiner	Fisher	http://discover.nci.nih.gov/gominer/ index.jsp



Enrichment Analysis: Factors to Consider

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





DAVID

	0	DAVID: Functional Annotation Tools	s - Mozilla Firefox
	Ē	jile <u>E</u> dit <u>V</u> iew Hi <u>s</u> tory <u>B</u> ookmarks <u>T</u> o	pols <u>H</u> elp
	1	DAVID: Functional Annotation Tools	+
	(A david.abcc.ncifcrf.gov/tools.js	sp
	Q	Most Visited 🗍 BaRC: Bioinformatics 🕻	📲 BLAST Help 🌹 Tak 2.0 🌠 BaRC Wiki 👫 EMBOSS WIBR 👙 String counter 🔗 NCBI 🦹 Perl mod 📓 UCSC Genome Browse 👫 BaRC analysis to
		DAVED	Analysis Wizard DAVID Bioinformatics Resources 6.7, NIAID/NIH
		Home Start Analysis Shortcu	t to DAVID Tools Technical Center Downloads & APIs Term of Service Why DAVID? About Us
	C	Upload List Background	Analysis Wizard
		Upload Gene List <u>Demolist 1 Demolist 2</u>	<u>Tell us how you like the tool</u> <u>Contact us for questions</u>
Gene List		<u>Upload Help</u> Step 1: Enter Gene List	🗲 Step 1. Submit your gene list through left panel.
		A: Paste a list QPRT GALNT3 PRDX2 GCH1 Clear Or B:Choose From a File	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 1255_g_at 1294_at 1316_at
Choose		Browse	1310_at 1320_at 1405_i_at 1431_at 1438_at
laentifier		Step 2: Select Identifier	1487_{at} 1494 <u>f</u> at
	· _	OFFICIAL_GENE_SYMBOL	1598_g_at
Select List Type		Step 3: List Type Gene List O Background O	
	-	Step 4: Submit List Submit List	





DAVID: Output

Contract of the state o	DAVID: Functional Annotation Result Eile Edit View Higtory Bookmarks Tools	Summary - Mozilla Firefox s Help								
	M DAVID: Functional Annotation Result Summary	+								
Performance of the section of the	david.abcc.nciferf.gov/summary.jsp						🏫 रु 🧭 🚼 र ncbi d	avid		$\left \right\rangle$
		Fui RȚABASE DAVID BI	nctional ioinformatio	Annotation Tool s Resources 6.7, NIAID/NIH		Expand "	'Options" to	chan	ge	
Cancel List Manager Annotation Summary Results Day SDD Ro To Benjaminis Second List Manager Current Gene List List_s Current Gene Contains Cur	Upload List Background		er Dowind	anz « whi? Tistul ni zstairs — Milà D'Aath		paramet	ers: eg. p-va	lue co	orrectio	n
det to land reaction by over the company of the co	Gene List Manager	Annotation Summ	ary Res	ults	ein and Tool Manual	by FDR o	or Benjamini			
Inter Part Inde United Systems I	Select to limit annotations by one or	Current Gene List: List_1 Current Background: Homo	n saniens	22793 DAVID IDs						
Bester Syncars Contrast, up, 1 4-33 4-34	-Use All Species - - Homo sepiens(1709) Mus musculus(1596)	 Disease (1 selected) Functional_Categories (3 selected) Gene Ontology (3 selected) 	lected)		DAVID: Database for Annotati david.abcc.ncfcrf.gov/chartRepo	on, Visualization, and Integrate	ed Discovery (Laboratory of Immunopa	thogenesis and Bio	pinformatics (LIB); Na	ional Institut (
List Manager Use Date Date Date Date Date List Manager Comps.gr.2 44.8 10.7 Comps.gr.2 44.8 10.7 Comps.gr.2 44.8 10.7 Comps.gr.2 44.8 10.7 Comps.gr.2 10.8 Manager Comps.gr.2 Manager Manager Ma	Rattus norvegicus(1544) 💌 Select Species	GOTERM_BP_1 64.5	5% 14705	Chart		Curronman	DAVID Bioinformat	ics Resour	ces 6.7	
List Manuager Lab Control, List, a 44, 44, 44, 44, 44, 44, 44, 44, 44,		GOTERM_BP_2 60.3	3% 13750			DATABASE	National Institute of Allergy and I	nfectious Disea:	ses (NIAID), NIH	
Image: Sect 14 Section 14 Mark 1107 Image: Sect 14 Sect 14 Sect 10 Sect	List Manager <u>Help</u>	GOTERM_BP_3 54.6	6% 12442	Chart	/					
		GOTERM_BP_4 49.0	0% 11177	Chart	Functional Anno	tation Chart				
Image: Second	Select List to:	GOTERM_BP_5 46.0	0% 10477	Chart	Cumont Cono lastu List				<u>1</u>	lelp and Manual
Pennove Containe is orther, p. y 1 is	Use Rename	GOTERM_BP_ALL 64.5	5% 14707	Chart	Current Gene Hst: List Current Background: Ho	no sapiens				
Show Gene Lie Outres (C, 1 41.96 929 Chart Diplay: Fold Enrichment Benjamini FORE Fisher Exact LT,PH,PT # 0 Records Tool View Ummapred Ide Outres (C, 1 41.96 929 Chart Diplay: Fold Enrichment Benjamini FORE Fisher Exact LT,PH,PT # 0 Records Tool Outres (C, 1 41.96 929 Chart Diplay: Fold Enrichment Benjamini FORE Fisher Exact LT,PH,PT # 0 Records Tool Outres (C, 1 41.96 929 Chart Diplay: Fold Enrichment Benjamini FOR Fisher Exact LT,PH,PT # 0 Records Tool Outres (C, 1 41.96 920 Chart Diplay: Fold Enrichment Benjamini FOR Fisher Exact LT,PH,PT # 0 Records Tool Diplay: Fold Enrichment Benjamini Fore Konte Enrichment Benjamini Fore Konte Enrichment Benjamini Fore Konte Fore Konte Fore Konte Fore Konte Fore <	Remove Combine	GOTERM_BP_FAT 34.7	7% 12474	Chart	22793 DAVID IDS					
Correct (E, 2) 41.1% 999 Chart Intresholds: Curu (2 Exet: (0.1 View Unmagned Lid: Correct (E, 2) 41.1% 999 Chart Intresholds: Curu (2 Exet: (0.1 Intresholds: Intresholds: Intresholds: Intresholds: Intresholds: Intresholds: Curu (2 Exet: (0.1 Intresholds: Intreshol	Show Gene List	GOTERM_CC_1 41.8	8% 9529	Chart	Options	01				
United Unitable 20142 		GOTERM_CC_2 41.1	1% 9369	Chart	Thresholds: Count 2				1000	
 	View Unmapped Ids	GOTERM_CC_3 41.1	1% 9362	Chart	Display: Fold Envictime	ent 🗆 Bonterroni 💌 Benjar	NINI LIFUR LIFISNEPEXACT LL	г,РН,РТ # ОТКЕС	oras 1000	
 		GOTERM_CC_4 37.2	2% 8469	Chart	Rerun Using Options Crea	te Sublist				
Sublet Catagory Iarm NI Genes Court X + Parameter Context, Cr, AT 1.8% 952 Chart 28.3% 666 Chart 226.31 Context, Mr, 1 73.1% 1652 Chart Gotext, Mr, 2 Diddoreductase activity RT 482 2.1 3.4E-314 2.2E-312 Context, Mr, 3 71.1% 16532 Chart Gotext, Mr, 2 transferses activity RT 482 2.1 3.4E-314 2.2E-312 Context, Mr, 3 71.1% 16392 Chart Gotext, Mr, 2 transferses activity RT 139 0.9 2.4E-136 5.3E-135 Context, Mr, 3 77.7% 13145 Chart Gotext, Mr, 2 transferses activity RT 139 0.9 2.4E-136 5.3E-135 Context, Mr, 3 77.7% 13145 Chart Gotext, Mr, 2 transferses activity RT 139 0.9 2.4E-13 1.5E-53 Context, Mr, 4 73.1% 16692 Chart Gotext, Mr, 2 transferses activity RT 126 2.6E-41 2.8E-13 <td></td> <td>GOTERM_CC_5 36.9</td> <td>9% 8420</td> <td>Chart</td> <td>25 chart records</td> <td></td> <td></td> <td></td> <td>6</td> <td>Download Fil</td>		GOTERM_CC_5 36.9	9% 8420	Chart	25 chart records				6	Download Fil
Pointen, ref, ar 0 29.3% 6666 Char		GOTERM_CC_ALL 41.8	8% 9529	Chart	Sublist <u>Category</u>	¢	<u>Term</u>	Genes <u>Cour</u>	nt 💠 <u>%</u> 💠 <u>P-Value</u>	Benjamini
Contenting 1 73.1% 16552 Charl 0 OOTERM_MF_2 transferase activity RT 597 2.6 5.2E-170 1.7E-168 Contenting 2 71.1% 1558 Charl 0 OOTERM_MF_2 ordedor binding RT 195 0.9 2.4E-136 5.3E-135 Contenting 3 61.1 5220 Charl 0 OOTERM_MF_2 transferase activity RT 103 0.5 3.1E-74 5.1E-73 Contenting 6 6.43% 1457 Charl 0 OOTERM_MF_2 transferase activity RT 103 0.5 3.1E-74 5.1E-73 Contenting 6 6.43% 1457 Charl 0 OOTERM_MF_2 transferase activity RT 103 0.5 3.1E-74 5.1E-73 Contenting 6 5.7% 1345 Charl 0 OOTERM_MF_2 transferase activity RT 0 0.4 1.2E-54 1.5E-53 Contenting 6 5.9% 1385 Charl 0 OOTERM_MF_2 transferase activity RT 272 0.3 2.6E-10 2.4E-135		GOTERM_CC_FAT 29.3	3% 6686	Chart	GOTERM_MF_2	oxidoreductase activity	RT	482	2.1 3.4E-314	2.2E-312
Corrent, MF, 2 71.1% 199 Chart 5.3E-135 Corrent, MF, 3 57.14 11229 Chart Chart 5.3E-135 Corrent, MF, 3 57.14 11229 Chart Chart 5.3E-135 Corrent, MF, 3 57.14 1129 Chart Chart S.3E-135 Corrent, MF, 4.1 64.3% 14857 Chart Chart Corrent, MF, 2 Varanno binding RI 113 0.5 3.1E-74 5.15-73 Corrent, MF, 5 57.7% 13145 Chart Chart Corrent, MF, 2 Varanno binding RI 79 0.3 2.6E-14 2.9E-40 Corrent, MF, ALL 73.1% 16552 Chart Corrent, MF, 2 Varanno binding RI 72 0.3 2.9E-40 2.8E-41 2.8E-41 2.8E-41 2.8E-41 2.8E-43 2.8E-41 2.8E-43 2.8E-43 </td <td></td> <td>GOTERM_MF_1 73.1</td> <td>1% 16652</td> <td>Chart</td> <td>GOTERM_MF_2</td> <td>transferase activity</td> <td><u>RT</u></td> <td>597</td> <td>2.6 5.2E-170</td> <td>1.7E-168</td>		GOTERM_MF_1 73.1	1% 16652	Chart	GOTERM_MF_2	transferase activity	<u>RT</u>	597	2.6 5.2E-170	1.7E-168
COTERNINE 3 5101 1229 Chat Chat S1E-73 S1E-73 Coternine MF, S1 643% 14857 Chat GOTERNINE, 2 Intransition Mining RI 90 0.4 1.2E-54 1.5E-53 Coternine MF, S1 57.7% 13145 Chat GOTERNINE, 2 Variancia binding RI 90 0.4 1.2E-54 1.5E-53 Coternine MF, S1 57.7% 13145 Chat GOTERNINE, 2 Variancia binding RI 79 0.3 2.6E-41 2.9E-40 Coternine MF, S1 57.9% 15925 Chat GOTERNINE, 2 Variancia binding RI 72 0.3 2.9E-30 2.3E-29 Coternine MF, FAT 59.9% 15925 Chat GOTERNINE, 2 carboxylic acid binding RI 72 0.3 2.9E-30 2.3E-29 Panther, PANTHER, BP_ALL 18.6% 4339 Chat GOTERNINE, 2 carboxylic acid binding RI 28 1.2 4.3E-16 2.9E-15 GOTERNINE, P_ALL 18.6% 4239 Chat GOTERNINE, 2 ocrease activity RI <td></td> <td>GOTERM_MF_2 71.1</td> <td>1% 16198</td> <td>Chart</td> <td>GOTERM_MF_2</td> <td>cofactor binding</td> <td><u>RT</u></td> <td>195</td> <td>0.9 2.4E-136</td> <td>5.3E-135</td>		GOTERM_MF_2 71.1	1% 16198	Chart	GOTERM_MF_2	cofactor binding	<u>RT</u>	195	0.9 2.4E-136	5.3E-135
Corrend Luft 64.3% 14857 Chart Cha		GOTERM_MF_3 67.1	15290	Chart	GOTERM_MF_2	lyase activity	<u>RT</u>	113	0.5 3.1E-74	5.1E-73
COT GOTERM_MF_S 57.7% 13145 Chart Chart 2.9E-40 abase: egs. COTERM_MF_ALL 73.1% 16552 Chart Chart CoteRM_MF_2 vitamin binding RI 79 0.3 2.6E-41 2.9E-40 abase: egs. CoteRM_MF_ALL 73.1% 16552 Chart CoteRM_MF_2 carbox/lic acid binding RI 72 0.3 2.9E-40 CoteRM_MF_BALL 59.% 15903 Chart CoteRM_MF_2 carbox/lic acid binding RI 72 0.3 2.9E-40 2.9E-40 Panther, PANTHER.BP_ALL 15.8% 133 Chart Chart CoteRM_MF_2 carbox/lic acid binding RI 2.9E-40 2.9E-40 GOTERM_MF_B_ALL 18.6% 433 Chart CoteRM_MF_2 carbox/lic acid binding RI 2.4E-15 2.9E-15 GOTERM_MF_2 0.0COTERM_MF_2		GOTERM ME 64.3	3% 14657	Chart	GOTERM_MF_2	tetrapyrrole binding	<u>RT</u>	90	0.4 1.2E-54	1.5E-53
abase: eg. Image: contract_wif_Ail 73.1% 16652 Chart Chart Contract_wif_Ail Chart Contract_wif_Ail Chart Contract_wif_Ail Contract_wif_Ail Chart Contract_wif_Ail	ect	GOTERM_MF_5 57.7	7% 13145	Chart	GOTERM_MF_2	vitamin binding	<u>RT</u>	· 79	0.3 2.6E-41	2.9E-40
abase: eg., Image: contract wife fail in the second multimate in the second multin the second multimate in the second multimate		GOTERM_MF_ALL 73.1	1% 16652	Chart	GOTERM_MF_2	nyaroiase activity	RT	 456 70 	2.0 3.UE-41	2.86-40
Panther, Plantiter, BPALL 18.0% 4131 Chart C	abase: eg.	GOTERM_MF_FAT	9% 15026	Chart	GOTERM MF 2	isomerase activity	RI DT	52	0.2 3.6E-16	2.4E-15
Panther, to expand Balancetsions (e selected) Balancetsions (e selected) Ba		PANTHER_BP_ALL 18.0	0% 4113	Chart	GOTERM MF 2	nucleoside binding	RT	284	1.2 4.3E-16	2.9E-15
to expand B Annotations (0 selected) B Main_Accessions (0 selected) B Main_Accessions (0 selected) B Athmax(0 sele	Panther.	PANTHER_MF_ALL 18.6	6% 4230	Chart	GOTERM_MF_2	oxygen binding	RI	27	0.1 1.0E-14	6.0E-14
to expand Exterior (# selected) B Main_Accessions (# selected) B Dathward (# construct)		General Annotations (0 select	ted)		GOTERM_MF_2	peroxidase activity	RT	23	0.1 2.3E-14	1.2E-13
a roundys (s selected)	to expand	 Literature (0 selected) Main_Accessions (0 selected) Pathways (3 selected))							





Gene Set Enrichment Analysis (GSEA)









Ranked by Enrichment Score (ES)





GSEA: Interpreting Results

(Je (Sk Yev Higtory (poknaks 3job))] Details for gene set (ESG, JAK, SYAT, SKIN. +



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





BiNGO

Cutanana Daulitan (Nau Saratan)		50		
File Edit View Select Layor: Plugins help				
	Search:			
Control Panel				
S Network VizMapper™ Editor Filters				
Network Nodes Edges				
		BiNGO Settings		
		BiNGO settings		
		Save settings as default	Help	
		Cluster name:		
		Get Cluster from Network Messte Genes from Text		
	Data Panel			
			×	
	lest: Hypergeometric	Do you want to assess over- or underrepresentation:		
	or Binomial	Overrepresentation Underrepresentation		
		Select a statistical test:		
	p-value correction	Hypergeometric test	✓	
		Select a multiple testing correction: Benjamini & Hochberg False Discovery Rate (FDR) correction	- Re	ef. or background.
		Choose a significance level:		cluding custom list
		0.05 Select the categories to be visualized:		
		Overrepresented categories after correction		Carra Orabala av
		Select reference set: Use whole annotation as reference set		Jene Untology
		Select ontology file:	Ľ	
		Select namespace:		
		biological_process Select organism/appotation:	~	
		Homo Sapiens	♥	
		Discard the following evidence codes:		
		Check box for saving Data Save BiNGO Data file in :		
		Start BiNGO		
				20
	Node Attribute Browser Edge Attribute Browser Network Attribute	Browser		
Welcome to Cytoscape 2.7.0 Right-click + drag to ZOOM	1 Middle-click + drag to PAN			





BiNGO: Output







GeneGO



- GO Processes
- GO Molecular Functions

Hide Description



-log(pValue)

- 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







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	а	С	а+с
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	
СоА	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:

FA cycle FA cycle Metabolic

		Metabolie
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)</pre>
- > 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?

