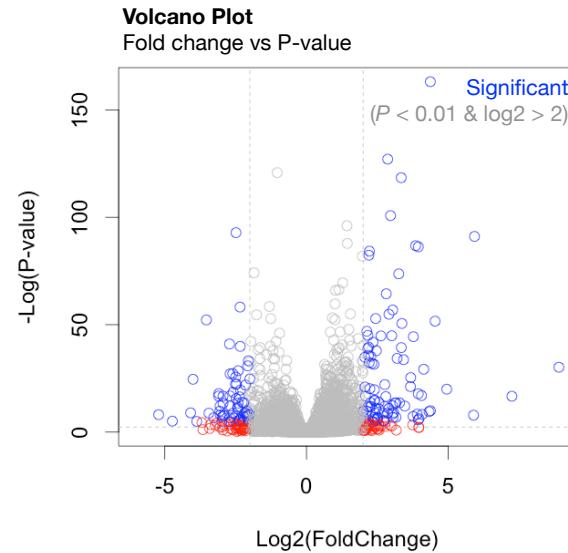


X	baseMean	log2FoldChange	IfSe	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARC1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	2.927283e-53	ARHGEF2
ENSG00000189221	2383.75371	3.3415441	0.21241508	15.731200	9.244206e-56	3.500088e-52	MAOA
ENSG00000120129	3440.70375	2.9652108	0.20370277	14.556557	5.306416e-48	1.607313e-44	DUSP1
ENSG00000148175	13493.92037	1.4271683	0.10036663	14.219550	6.929711e-46	1.749175e-42	STOM
ENSG00000178695	2685.40974	-2.4890689	0.17806407	-13.978501	2.108817e-44	4.562576e-41	KCTD12
ENSG00000109906	439.54152	5.9275950	0.42819442	13.843233	1.397758e-43	2.646131e-40	ZBTB16
ENSG00000134686	2933.64246	1.4394898	0.10582729	13.602255	3.882769e-42	6.533838e-39	PHC2
ENSG00000101347	14134.99177	3.8504143	0.28490701	13.514635	1.281894e-41	1.941428e-38	SAMHD1
ENSG00000096060	2630.23049	3.9450524	0.29291821	13.468102	2.409807e-41	3.317866e-38	FKBP5
ENSG00000166741	7542.25287	2.2195906	0.16673544	13.312050	1.970000e-40	2.486304e-37	NNMT
ENSG00000125148	3695.87946	2.1985636	0.16700546	13.164621	1.402400e-39	1.633797e-36	MT2A
ENSG00000162614	5646.18314	1.9711402	0.15020631	13.122885	2.434854e-39	2.633990e-36	NEXN
ENSG00000106976	989.04683	-1.8501713	0.14778657	-12.519211	5.861471e-36	5.918132e-33	DNM1
ENSG00000187193	199.07694	3.2551424	0.26090711	12.476250	1.006146e-35	9.523804e-33	MT1X
ENSG00000256235	1123.47954	1.2801193	0.10547438	12.136779	6.742862e-34	6.007096e-31	SMIM3
ENSG00000177666	2639.57020	1.1399947	0.09606884	11.866436	1.768422e-32	1.487930e-29	PNPLA2
ENSG00000164125	7257.00808	1.0248523	0.08657600	11.837603	2.494830e-32	1.988642e-29	FAM198B
ENSG00000198624	2020.04495	2.8141014	0.24063429	11.694515	1.359615e-31	1.029569e-28	CCDC69
ENSG00000123562	5008.55294	1.0045453	0.08901501	11.285123	1.554241e-29	1.120904e-26	MORF4L2
ENSG00000144369	1283.77980	-1.3090041	0.11714863	-11.173875	5.473974e-29	3.768333e-26	FAM171B
ENSG00000196517	241.91536	-2.3456877	0.21047366	-11.144804	7.591120e-29	4.998588e-26	SLC6A9
ENSG00000135821	19973.40000	3.0413943	0.27601796	11.018828	3.100706e-28	1.956675e-25	GLUL



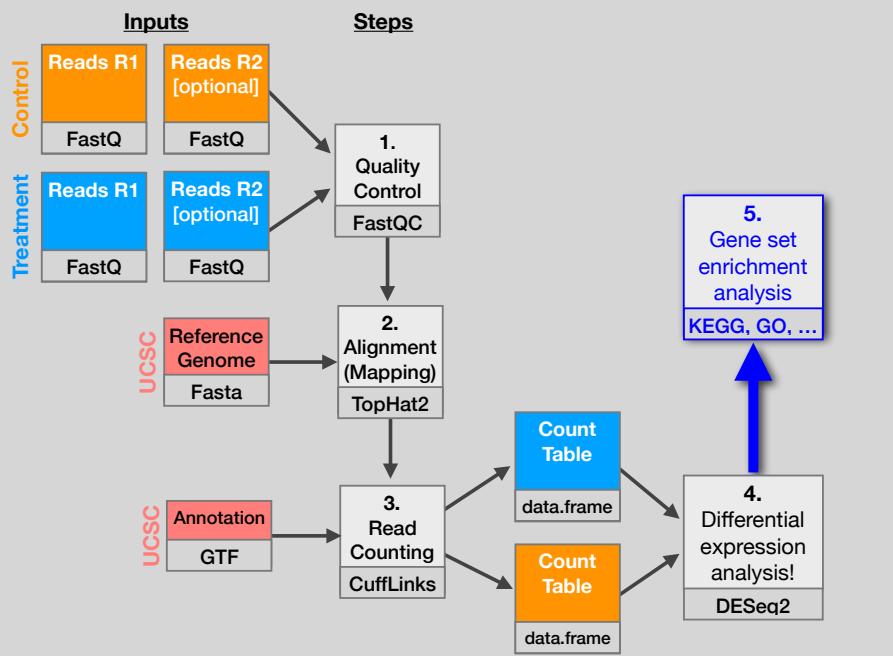
My high-throughput experiment generated a long list of genes/proteins...

What do I do now?



## Pathway analysis! (a.k.a. geneset enrichment)

Use bioinformatics methods to help extract biological meaning from such lists...



## Basic idea

Differentially Expressed Genes (DEGs)

ENSG00000152881	9147.2093	4.3481590	0.21715644	18.421286	8.697079e-76	L142028-71	SPARC1
ENSG00000179094	741.25248	2.4938590	0.17555853	13.613105	7.972621e-60	L072707-56	PMS1
ENSG00000116584	227.716345	-1.0149545	0.05005729	13.505357	5.798118e-57	S202788-53	ARHGAP2
ENSG00000186221	238.737573	3.1950544	0.21454507	13.511200	8.246000e-56	S300515-52	MAGA
ENSG00000177357	1.3952354	0.00000000	0.00000000	13.511200	8.246000e-56	S300515-52	MAGA
ENSG00000141175	1349.62037	1.4271683	0.10096663	14.121950	6.039711e-46	L740174-42	STOM
ENSG00000178695	2618.40074	-2.4906649	0.17060460	13.972651	7.008181e-44	L562724-41	KCTD12
ENSG00000109640	419.54152	5.9737950	0.41618442	13.641233	3.397578e-41	L264310-40	ZEBR1
ENSG00000110000	1.3952354	0.00000000	0.00000000	13.511200	8.246000e-56	S300515-52	MAGA
ENSG00000113147	14134.9037	3.4504143	0.24490701	13.514653	1.384084e-41	L341628-38	SAMHD1
ENSG00000137049	2610.23049	3.9450524	0.29591803	13.641010	2.608070e-41	S317864-38	FBXRS
ENSG00000164741	7542.23287	2.2159510	0.16473104	13.131050	1.970000e-40	L488304-37	NMNT
ENSG00000125148	360.57946	2.1895036	0.16170546	13.161400	1.602400e-39	L131797-36	MTZA
ENSG00000106076	1.3952354	0.00000000	0.00000000	13.511200	8.246000e-56	S300515-52	MAGA
ENSG00000106076	988.66843	-1.8501713	0.14798537	12.531811	8.814784e-36	S301310-33	DNAI1
ENSG00000187131	139.07694	3.2511424	0.26099711	12.478250	1.008464e-31	S323804-33	MTLX
ENSG00000262635	11234.9794	1.2801210	0.15047433	12.118979	6.742826e-34	L007706-31	SMN3
ENSG00000177668	2619.57020	1.1599497	0.09068884	13.886458	1.784022e-32	L437936-32	PPML2
ENSG00000188624	2020.04495	2.8161014	0.24681629	16.094515	1.559815e-31	L029598-28	CCDC89
ENSG00000123582	5098.555294	1.0045453	0.08001050	11.285123	1.559421e-29	L129504-26	MORF4L2
ENSG00000144569	1283.77980	-1.3050041	0.11714680	11.171875	5.473974e-29	S376338-26	FAM171B
ENSG00000196157	201.81536	-2.3456577	0.21097986	11.144001	7.951220e-29	S495858-26	SCEAS
ENSG00000137049	1692.46020	3.9451043	0.27601936	11.030023	4.107070e-21	L095072-21	C11B

Gene-sets (Pathways, annotations, etc...)



## Basic idea

### Differentially Expressed Genes (DEGs)

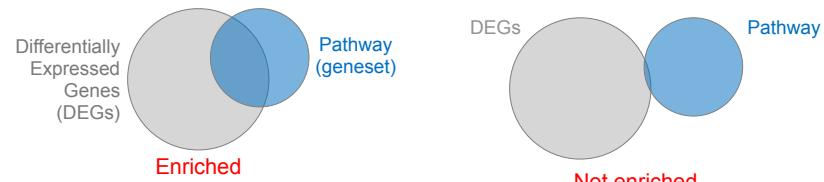
	baseMean	logFoldChange	negLog10PValue	padj	symbol
ENSG0000011583	954.77093	3.463556	0.171134	0.002704	MAP3K1
ENSG0000017094	743.52000	2.830000	0.175142	0.002704	MAP3K1
ENSG0000016584	2277.81345	-1.034700	0.00305273	0.003557	MHCII2
ENSG0000018202	2383.73373	1.341554	0.21241508	0.003557	MHCII2
ENSG0000018019	3446.70373	2.961208	0.2070277	0.003557	OASPF1
ENSG0000014375	13493.8037	1.427748	0.100463	0.003557	STOM
ENSG0000014376	13493.8037	1.427748	0.100463	0.003557	STOM
ENSG0000013996	438.44142	5.927950	0.42818442	0.004111	ZEB1B
ENSG0000014246	2933.64246	1.439488	0.15827279	0.004225	PDCD2
ENSG0000013347	1414.99177	3.850541	0.24895703	0.004317	SAMO1
ENSG0000013348	2030.52000	3.850541	0.24895703	0.004317	SAMO1
ENSG0000014261	7642.52327	2.710898	0.14673544	0.004317	WNT7
ENSG0000011148	3095.87946	2.189888	0.17050546	0.004317	WT2A
ENSG0000012616	1646.18314	1.971140	0.15020831	0.004317	WTN1
ENSG0000014376	1732.77686	1.895173	0.15020831	0.004317	WTN1
ENSG0000018202	1121.47954	1.295119	0.15674748	0.004778	WTN1
ENSG0000017696	2639.37020	1.119994	0.09008684	0.004848	YAP1A2
ENSG0000014123	7257.20088	1.024852	0.08657600	0.004848	YAP1B
ENSG0000014064	2026.04495	2.814104	0.24895703	0.004848	ZCCHC4
ENSG0000014496	1281.77980	1.502000	0.15020831	0.004848	ZFAT2
ENSG0000014498	1281.77980	-1.359004	0.13714663	0.004848	ZFAT2
ENSG0000018517	241.85156	-2.345887	0.21047366	0.011400	ZFAT2
ENSG0000011582	19075.40029	1.041394	0.27021295	0.011400	ZIC4

### Gene-sets (Pathways, annotations, etc...)

Annotate...



## Pathway analysis (a.k.a. geneset enrichment) Principle



- DEGs come from your experiment

> Critical, needs to be as clean as possible

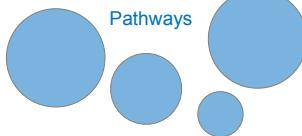
- Pathway genes ("geneset") come from annotations

> Important, but typically not a competitive advantage

- Variations of the math: overlap, ranking, networks... > Not critical, different algorithms show similar performances

### Differentially Expressed Genes (DEGs)

Overlap...



## Pathway analysis (a.k.a. geneset enrichment) Limitations

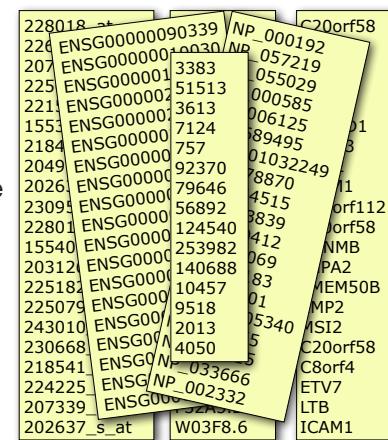
Side-note:

- **Geneset annotation bias:** can only discover what is already known
- **Non-model organisms:** no high-quality genesets available
- **Post-transcriptional regulation** is neglected
- **Tissue-specific** variations of pathways are not annotated
  - e.g. NF-κB regulates metabolism, not inflammation, in adipocytes
- **Size bias:** stats are influenced by the size of the pathway
  - Many pathways/receptors **converge** to few regulators
    - e.g. Tens of innate immune receptors activate four TFs: NF-κB, AP-1, IRF3/7, NFAT

## Starting point for pathway analysis: Your gene list

- You have a list of genes/proteins of interest
- You have quantitative data for each gene/protein

- Fold change
- p-value
- Spectral counts
- Presence/absence



## Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id

## Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- **Various web sites translate ids -> best for small lists**
  - UniProt <[www.uniprot.org](http://www.uniprot.org)>; IDConverter <[idconverter.bioinfo.cnio.es](http://idconverter.bioinfo.cnio.es)>

## Translating between identifiers: UniProt <[www.uniprot.org](http://www.uniprot.org)>

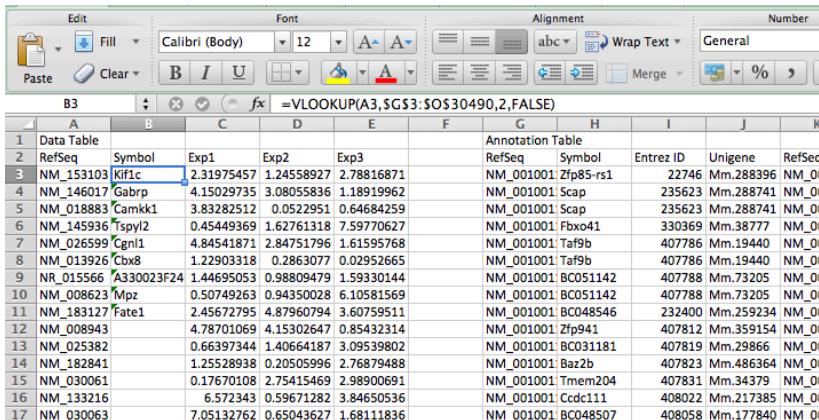
The screenshot shows the UniProt homepage. At the top, there is a search bar with dropdown menus for 'Search in' (Protein Knowledgebase (UniProtKB)) and 'Query'. Below the search bar are buttons for 'Search', 'Clear', 'Fields', 'Blast', 'Align', 'Retrieve', and 'ID Mapping'. The 'ID Mapping' button is highlighted with a red box. The main area is titled 'WELCOME' and 'NEWS'. Below this, there is a form for 'Identifiers' mapping. It has two dropdown menus: 'From' (set to 'EMBL/GenBank/DDBJ') and 'To' (set to 'UniProtKB AC'). There is also a file upload field labeled 'Choose File' with the placeholder 'no file selected'. To the right of the dropdowns are three buttons: 'Map', 'Swap', and 'Clear'.

## Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> *best for small lists*
  - UniProt <[www.uniprot.org](http://www.uniprot.org)>; IDConverter <[idconverter.bioinfo.cnio.es](http://idconverter.bioinfo.cnio.es)>
- **VLOOKUP in Excel - good if you are an excel whizz - I am not!**
  - Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the 2 IDs you want to convert between; Sort by ID; Use vlookup to translate your list

## Translating between identifiers: Excel VLOOKUP

VLOOKUP(lookup\_value, table\_array, col\_index\_num)



The screenshot shows an Excel spreadsheet with two tables. The first table, 'Data Table', has columns A through F. The second table, 'Annotation Table', has columns G through K. Cell B3 contains the formula =VLOOKUP(A3,\$G\$3:\$O\$30490,2, FALSE). The 'Annotation Table' is sorted by RefSeq.

A	B	C	D	E	F	G	H	I	J	K
RefSeq	Symbol	Exp1	Exp2	Exp3		RefSeq	Entrez ID	Unigene	RefSeq	
NM_153103	Kif1c	2.31975457	1.24558927	2.78816871		NM_001001	Zfp85-rs1	22746	Mm.288396	NM_001
NM_146017	Gabbrp	4.15029735	3.08055836	1.18919962		NM_001001	Scap	235623	Mm.288741	NM_001
NM_018883	Camkk1	3.83282512	0.0522951	0.64684259		NM_001001	Scap	235623	Mm.288741	NM_001
NM_145936	Tspyl2	0.45449369	1.62761318	7.59770627		NM_001001	Fbxo41	330369	Mm.38777	NM_001
NM_026599	Cgnl1	4.84541871	2.84751796	1.61595768		NM_001001	Taf9b	407786	Mm.19440	NM_001
NM_013926	Cbx8	1.22903318	0.2863077	0.02952665		NM_001001	Taf9b	407786	Mm.19440	NM_001
NR_015566	A330023F24	1.44695053	0.98809479	1.59330144		NM_001001	BC051142	407788	Mm.73205	NM_001
NM_008623	Mpz	0.50749263	0.94350028	6.10581569		NM_001001	BC051142	407788	Mm.73205	NM_001
NM_183127	Fate1	2.45672795	4.87960794	3.60759511		NM_001001	BC048546	232408	Mm.259234	NM_001
NM_008943		4.78701069	4.15302647	0.85432314		NM_001001	Zfp941	407812	Mm.359154	NM_001
NM_025382		0.66397344	1.40664187	3.09539802		NM_001001	BC031181	407819	Mm.29866	NM_001
NM_182841		1.25528938	0.20505996	2.76879488		NM_001001	Baz2b	407823	Mm.486364	NM_001
NM_030061		0.17670108	2.75415469	2.98900691		NM_001001	Tmem204	407831	Mm.34379	NM_001
NM_133216		6.572343	0.59671282	3.84650536		NM_001001	Ccdc111	408022	Mm.217385	NM_001
NM_030063		7.05132762	0.65043627	1.68111836		NM_001001	BC048507	408058	Mm.177840	NM_001

## Translating between identifiers

- Many different identifiers exist for genes and proteins, e.g. UniProt, Entrez, etc.
- Often you will have to translate one set of ids into another
  - A program might only accept certain types of ids
  - You might have a list of genes with one type of id and info for genes with another type of id
- Various web sites translate ids -> best for small lists
  - UniProt < [www.uniprot.org](http://www.uniprot.org); IDConverter < [idconverter.bioinfo.cnio.es](http://idconverter.bioinfo.cnio.es)>
- VLOOKUP in Excel -> good if you are an excel whizz - I am not!
  - Download flat file from Entrez, Uniprot, etc; Open in Excel; Find columns that correspond to the two ids you want to convert between; Use vlookup to translate your list

- Use the **merge()** or **mapIDs()** functions in R - fast, versatile & reproducible!
  - Also **clusterProfiler::bitr()** function and many others... [Link to clusterProfiler vignette]

Reminder

2. class-material (bash)

```
# Using the merge() function
> anno <- read.csv("data/annotables_grch38.csv") This is an annotation file

> merge(mygenes, anno, by.x="row.names", by.y= "ensgene")
This is our differential expressed genes
```

Reminder

2. class-material (bash)

```
# Using the merge() function
> anno <- read.csv("data/annotables_grch38.csv")

> merge(mygenes, anno, by.x="row.names", by.y= "ensgene")

# Using mapIDs() function from bioconductor
> library("AnnotationDbi")
> library("org.Hs.eg.db") Load the required Bioconductor packages

> mygenes$symbol <- mapIds( org.Hs.eg.db,
  column="SYMBOL",
  keys=row.names(mygenes),
  keytype="ENSEMBL") Annotation we want to add
  Our vector of gene names & their format
```

## bitr: Biological Id Translator

clusterProfiler provides `bitr` and `bitr_kegg` for converting ID types. Both `bitr` and `bitr_kegg` support many species including model and many non-model organisms.

```
x <- c("GPX3", "GLRX", "LBP", "CRYAB", "DEFB1", "HCLS1", "SOD2", "HSPA2",  
      "ORM1", "IGFBP1", "PTHLH", "GPC3", "IGFBP3", "T0B1", "MITF", "NDRG1",  
      "NR1H4", "FGFR3", "PVR", "IL6", "PTPRM", "ERBB2", "NID2", "LAMB1",  
      "COMP", "PLS3", "MCAM", "SPP1", "LAMC1", "COL4A2", "COL4A1", "MYOC",  
      "ANXA4", "TFPI2", "CST6", "SLPI", "TIMP2", "CPM", "GGT1", "NNMT",  
      "MAL", "EEF1A2", "HGD", "TCN2", "CDA", "PCCA", "CRYM", "PDHX",  
      "STC1", "WARS", "HMOX1", "FXYD2", "RBP4", "SLC6A12", "KDELR3", "ITM2B")  
eg = bitr(x, fromType="SYMBOL", toType="ENTREZID", OrgDb="org.Hs.eg.db")  
head(eg)
```

```
## SYMBOL ENTREZID  
## 1 GPX3 2878  
## 2 GLRX 2745  
## 3 LBP 3929  
## 4 CRYAB 1410  
## 5 DEFB1 1672  
## 6 HCLS1 3059
```

See package vignette:

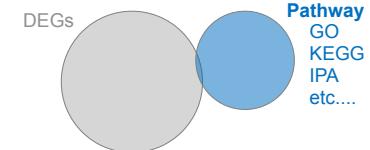
<https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html>

Alternative...

## What functional set databases do you want?

- Most commonly used:

- **Gene Ontology (GO)**
- **KEGG Pathways** (mostly metabolic)
- **GeneGO MetaBase** 
- **Ingenuity Pathway Analysis (IPA)** 



- Many others...

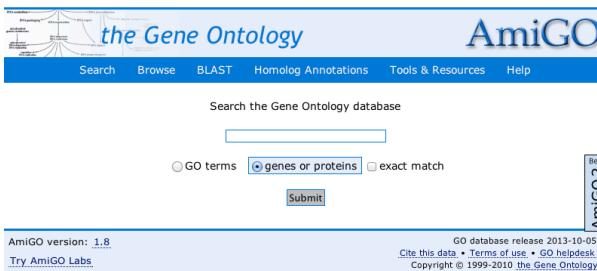
- **Enzyme Classification, PFAM, Reactome,**
- Disease Ontology, MSigDB, Chemical Entities of Biological Interest, Network of Cancer Genes etc...
- See: Open Biomedical Ontologies ([www.obofoundry.org](http://www.obofoundry.org))

## GO < [www.geneontology.org](http://www.geneontology.org) >

- What function does HSF1 perform?
  - response to heat; sequence-specific DNA binding; transcription; etc
- **Ontology** => a structured and controlled vocabulary that allows us to annotate gene products consistently, interpret the relationships among annotations, and can easily be *handled by a computer*
- GO database consists of 3 ontologies that describe gene products in terms of their associated **biological processes**, **cellular components** and **molecular functions**

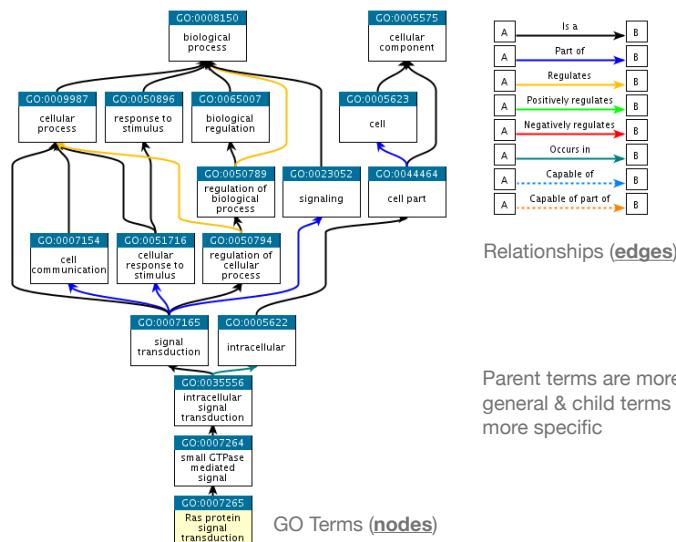
## GO Annotations

- GO is not a stand-alone database of genes/proteins or sequences
- Rather gene products get annotated with **GO terms** by UniProt and other organism specific databases, such as Flybase, Wormbase, MGI, ZFIN, etc.
- Annotations are available through AmiGO < [amigo.geneontology.org](http://amigo.geneontology.org) >



The screenshot shows the AmiGO web interface. At the top, there's a navigation bar with links for 'Search', 'Browse', 'BLAST', 'Homolog Annotations', 'Tools & Resources', and 'Help'. Below the navigation bar is a search bar labeled 'Search the Gene Ontology database'. Underneath the search bar are three radio buttons: 'GO terms' (selected), 'genes or proteins', and 'exact match'. A 'Submit' button is located below the search bar. At the bottom of the page, there's footer information including 'AmiGO version: 1.8', 'GO database release 2013-10-05', 'Try AmiGO Labs', 'Cite this data', 'Terms of use', 'GO Helpdesk', 'Copyright © 1999-2010 the Gene Ontology', and a 'Beta' badge.

**GO is structured as a “directed graph”**

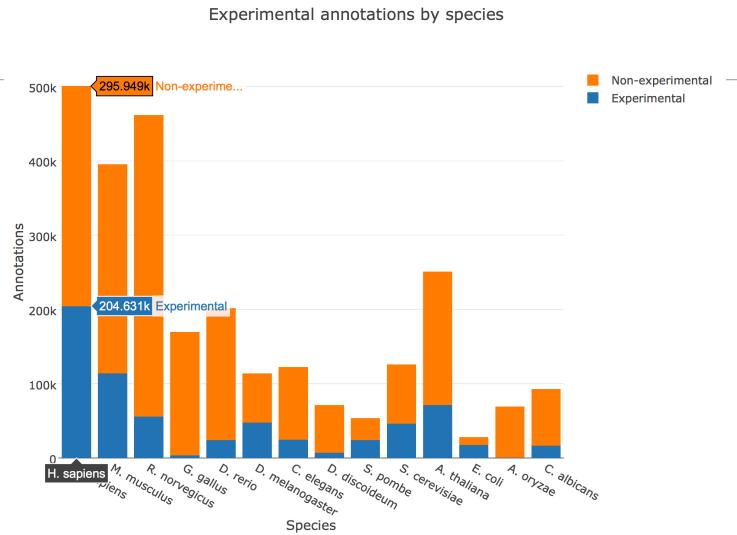


Parent terms are more general & child terms more specific

Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from mutant phenotype	Experimental	Yes	61,549
IPI	Inferred from physical interaction	Experimental	Yes	17,043
ISS	Inferred from sequence or structural similarity	Computational	Yes	196,643
RCA	Inferred from reviewed computational analysis	Computational	Yes	103,792
IGC	Inferred from genomic context	Computational	Yes	4
IEA	Inferred from electronic annotation	Computational	No	15,687,382
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes	5,167
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes	44,564
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes	25,656
ND	No biological data available	No information available	Yes	132,192
NR	Not recorded	Unknown	Yes	1,185

\*October 2007 release

Can now do gene list analysis with GeneGO online!



- See AmiGO for details: [http://amigo.geneontology.org/amigo/base\\_statistics](http://amigo.geneontology.org/amigo/base_statistics)

pantherdb.org/webservices/go/overrep.jsp

# PANTHER Classification System

GENEONTOLOGY Unifying Biology

**New! PANTHER13.1 released.**

**Gene List Analysis**   **Browse**   **Sequence Search**   **cSNP Scoring**   **Keyword Search**

**Search**

All

**Go**

**Quick links**

Whole genome function views  
Genome statistics  
Data Version  
How to cite PANTHER  
**NEW!** Recent publication describing PANTHER

**News**

PANTHER13.1 Released  
Click for additional info.

**Newsletter subscription**

Enter your Email:

 PostgreSQL

Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page.

*Error parsing request, no input specified*

**Help Tips**

**Steps:**

- > 1. Select list and list type to analyze
- > 2. Select Organism
- > 3. Select operation

**1.** Enter IDs:  Supported ID types  
Enter IDs and/or select file for batch upload. Else enter IDs or select file or list from workspace for comparing to a reference list.  
 separate IDs by a space or comma  
 no file selected

Please login to be able to select lists from your workspace.

**Select List Type:**

ID List  
 Previously exported text search results  
 Workspace list  
 PANTHER Generic Mapping File  
 VCF File Flanking region 20 Kb

**2.** Select organism.  
 Homo sapiens  
 Mus musculus  
 Rattus norvegicus  
 Gallus gallus  
 Danio rerio

**3.** Select analysis.  
 Functional classification viewed in gene list

## Another popular online tool: **DAVID** at NIAID < david.abcc.ncifcrf.gov >

The screenshot shows the DAVID Analysis Wizard interface. At the top, there's a navigation bar with links like Home, Start Analysis, Shortcut to DAVID Tools, Technical Center, Downloads & APIs, Term of Service, Why DAVID?, and About Us. Below the navigation bar, there's a section titled "Analysis Wizard" with the subtitle "DAVID Bioinformatics Resources 2008, NIAID/NIH". The main area is titled "Analysis Wizard" and contains four steps:

- Step 1: Enter Gene List**: A text input field with a "Clear" button. A note says: "Note: Affy Exon IDs and Affy Gene Array IDs are now supported in DAVID, as \*affy\_id\* type." An example list is provided: 1007\_s\_at, 1008\_s\_at, 117\_at, 121\_at, 1255\_g\_at, 1316\_at, 1320\_at, 1431\_at, 1438\_at, 1487\_at, 1494\_f\_at, 1598\_g\_at.
- Step 2: Select Identifier**: A dropdown menu set to "AFFY\_ID".
- Step 3: List Type**: Radio buttons for "Gene List" (selected) and "Background".
- Step 4: Submit List**: A "Submit List" button.

## DAVID

- Functional Annotation Chart



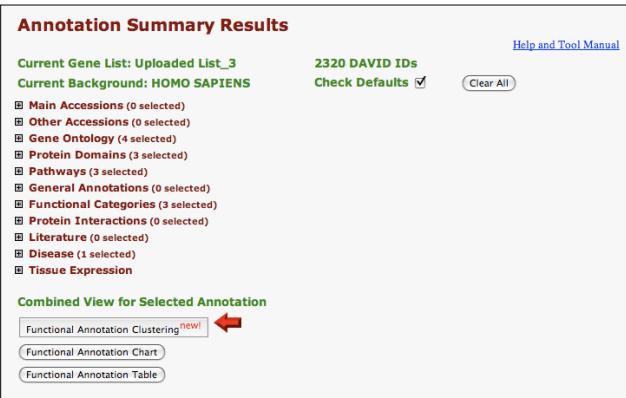
Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources  
Da Wei Huang, Brad T Sherman & Richard A Lempicki  
Nature Protocols 4, 44 - 57 (2009)

## Overlapping functional sets

- Many functional sets overlap
  - In particular those from databases that are hierarchical in nature (e.g. GO)
- Hierarchy enables:
  - Annotation flexibility (e.g. allow different degrees of annotation completeness based on what is known)
  - Computational methods to “understand” function relationships (e.g. ATPase function is a subset of enzyme function)
- Unfortunately, this also makes functional profiling trickier
  - Clustering of functional sets can be helpful in these cases

## DAVID

- DAVID now offers functional annotation clustering:



## DAVID Functional Annotation Clustering

- Based on shared genes between functional sets



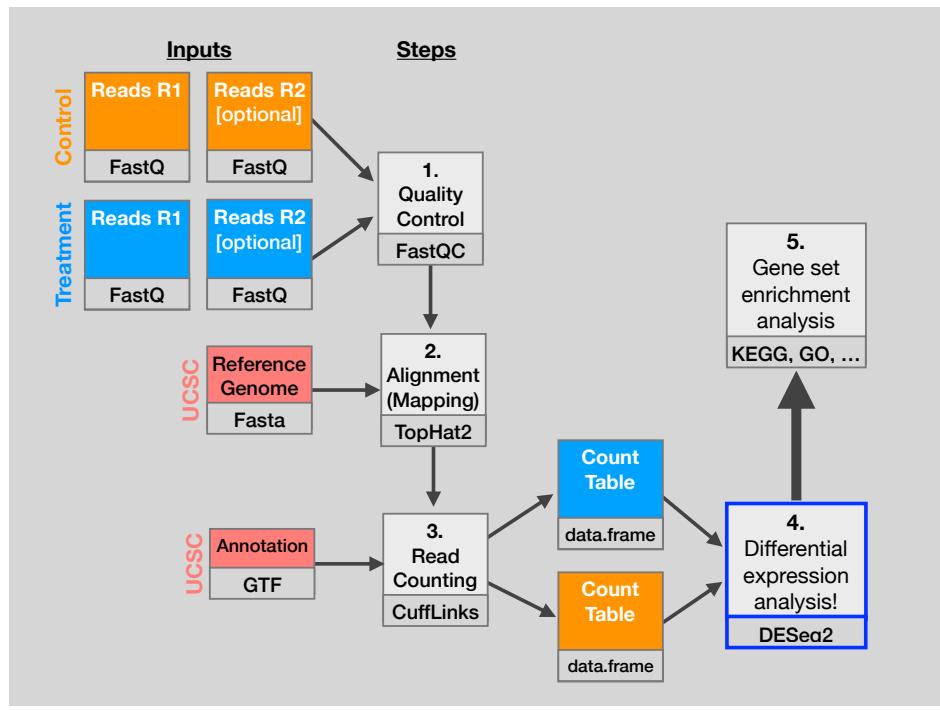
## Hands-on time!

<http://thegrantlab.org/bggn213>

Want more?



- GeneGO** < [portal.genego.com](http://portal.genego.com) >
  - MD/PhD curated annotations, great for certain domains (eg, Cystic Fibrosis)
  - Nice network analysis tools
  - Email us for access
- Oncomine** < [www.oncomine.org](http://www.oncomine.org) >
  - Extensive cancer related expression datasets
  - Nice concept analysis tools
  - Research edition is free for academics, Premium edition \$\$\$
- Lots and lots other R/Bioconductor packages in this area!!!**



# counts + metadata

## 1 countData

gene	ctrl_1	ctrl_2	exp_1	exp_2
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
...	...	...	...	...

countData is the count matrix  
(Number of reads coming from each gene for each sample)

N.B. First column of colData must match column names (i.e. sample names) of countData (-1st)

## 2 colData

id	treatment	sex	...
ctrl_1	control	male	...
ctrl_2	control	female	...
exp_1	treated	male	...
exp_2	treated	female	...

colData describes metadata about the columns of countData

## Inputs

Reads R1  
FastQ

Reads R2 [optional]  
FastQ

Reads R1  
FastQ

Reads R2 [optional]  
FastQ

Reference Genome  
Fasta

Annotation  
GTF

## Steps

1. Quality Control

FastQC

2. Alignment (Mapping)

TopHat2

3. Read Counting

CuffLinks

Count Table  
data.frame

Count Table  
data.frame

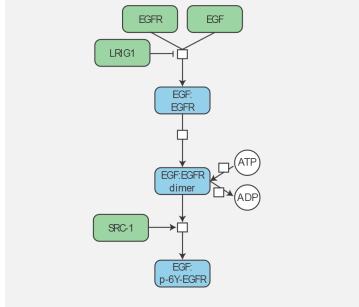
4. Differential expression analysis!

DESeq2

5. Gene set enrichment analysis  
KEGG, GO, ...

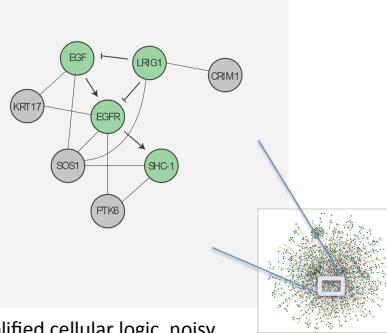
## Pathways vs Networks

EGFR-centered Pathway



- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

EGFR-centered Network



- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from omics data integration

Next Class

## Goal

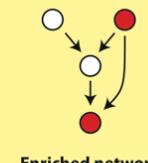
### 1 Enrichment of fixed gene sets

Identification of pre-built pathways or networks that are enriched in a set of mutated or differentially expressed genes

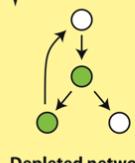
### 2 De novo sub-network construction and clustering

Construction of specific sub-networks from the set of mutated or differentially expressed genes to identify an extended list of putative cancer genes

## Output



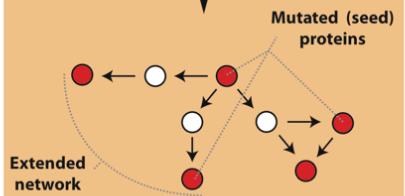
Enriched network



Depleted network

Next Class

Mutated (seed) proteins



Extended network

## Goal

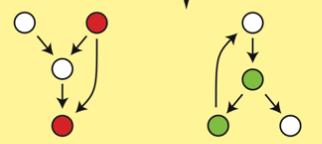
### 1 Enrichment of fixed gene sets

Identification of pre-built pathways or networks that are enriched in a set of mutated or differentially expressed genes

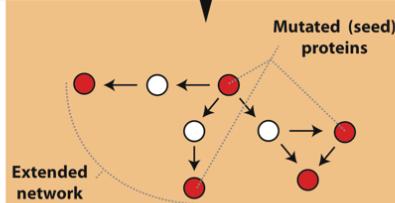
### 2 De novo sub-network construction and clustering

Construction of specific sub-networks from the set of mutated or differentially expressed genes to identify an extended list of putative cancer genes

## Output



What biological process is altered in this cancer?



Are NEW pathways altered in this cancer? Are there clinically relevant tumor subtypes?

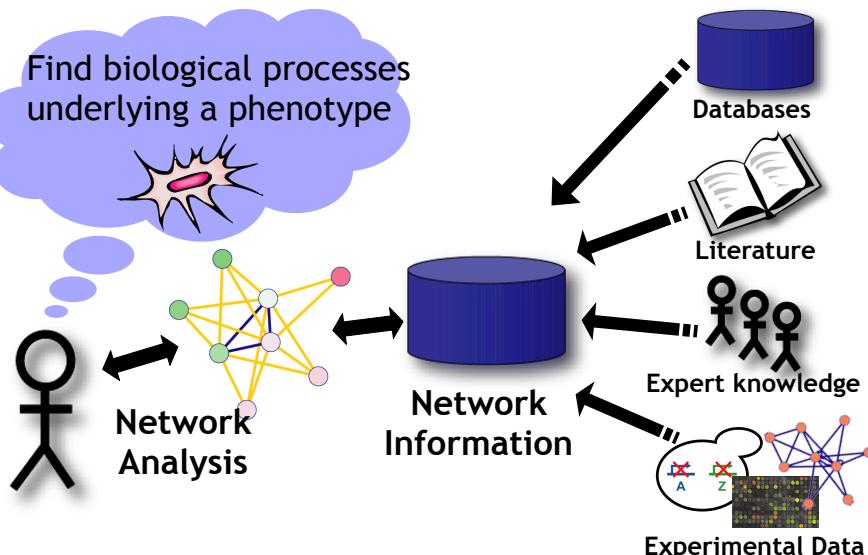
Next Class

Side-note:

## Pathway analysis (a.k.a. geneset enrichment) Limitations

- **Geneset annotation bias:** can only discover what is already known
- **Non-model organisms:** no high-quality genesets available
- **Post-transcriptional regulation** is neglected
- **Tissue-specific** variations of pathways are not annotated
  - e.g. NF- $\kappa$ B regulates metabolism, not inflammation, in adipocytes
- **Size bias:** stats are influenced by the size of the pathway
  - Many pathways/receptors **converge** to few regulators  
e.g. Tens of innate immune receptors activate four TFs:  
NF- $\kappa$ B, AP-1, IRF3/7, NFAT

## Pathway & Network Analysis Overview

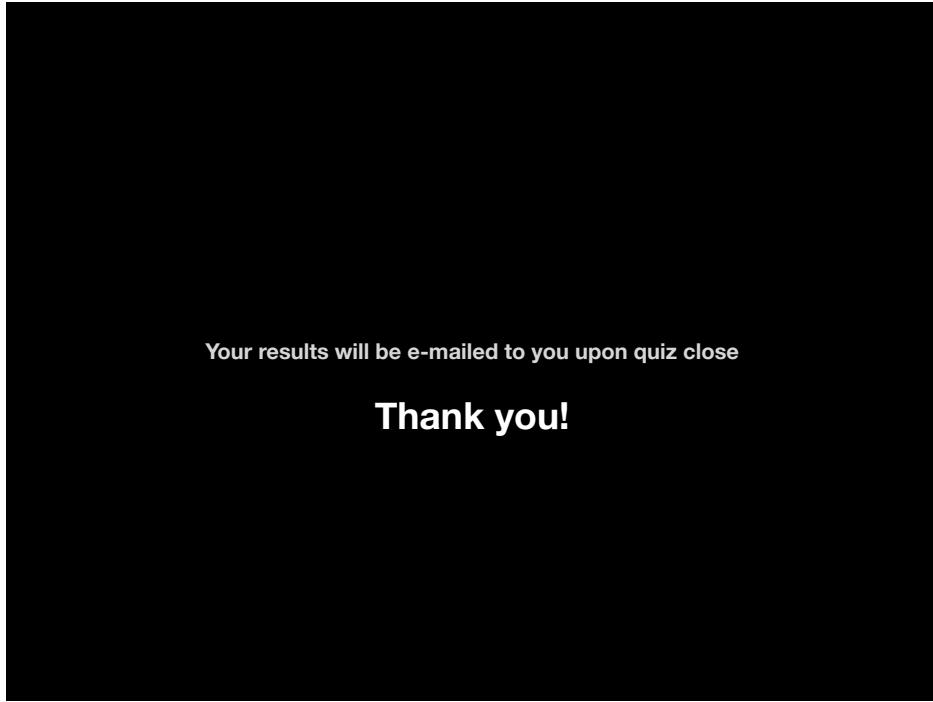


Do it Yourself!

## R Knowledge Check For BGGN-213 Quiz

This will be marked but not graded  
(i.e. will not factor into your course grade)

Time Limit: 30 mins



Your results will be e-mailed to you upon quiz close

**Thank you!**