

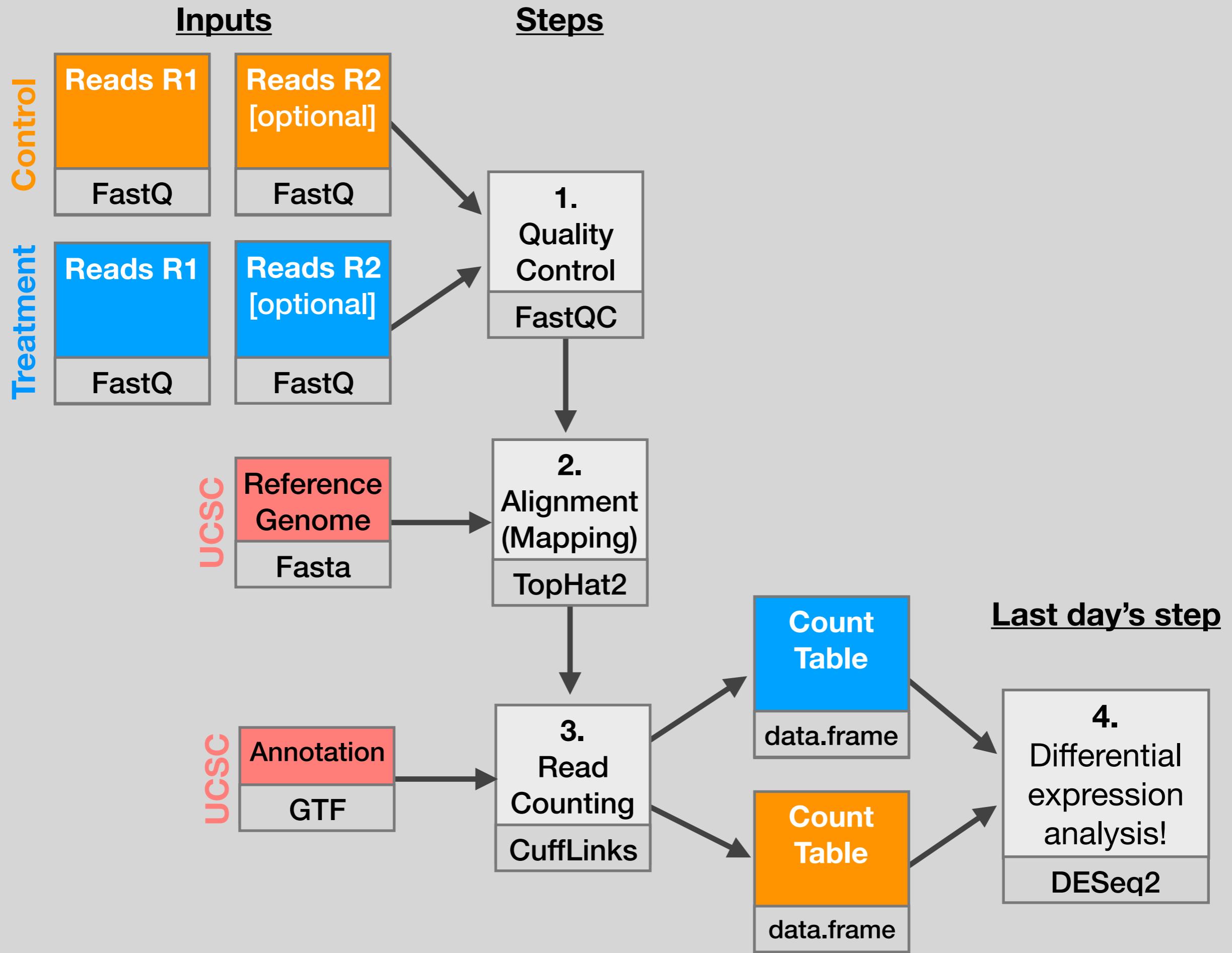
# BIMM 143

## Pathway Analysis and the Interpretation of Gene Lists

Lecture 15

Barry Grant  
UC San Diego

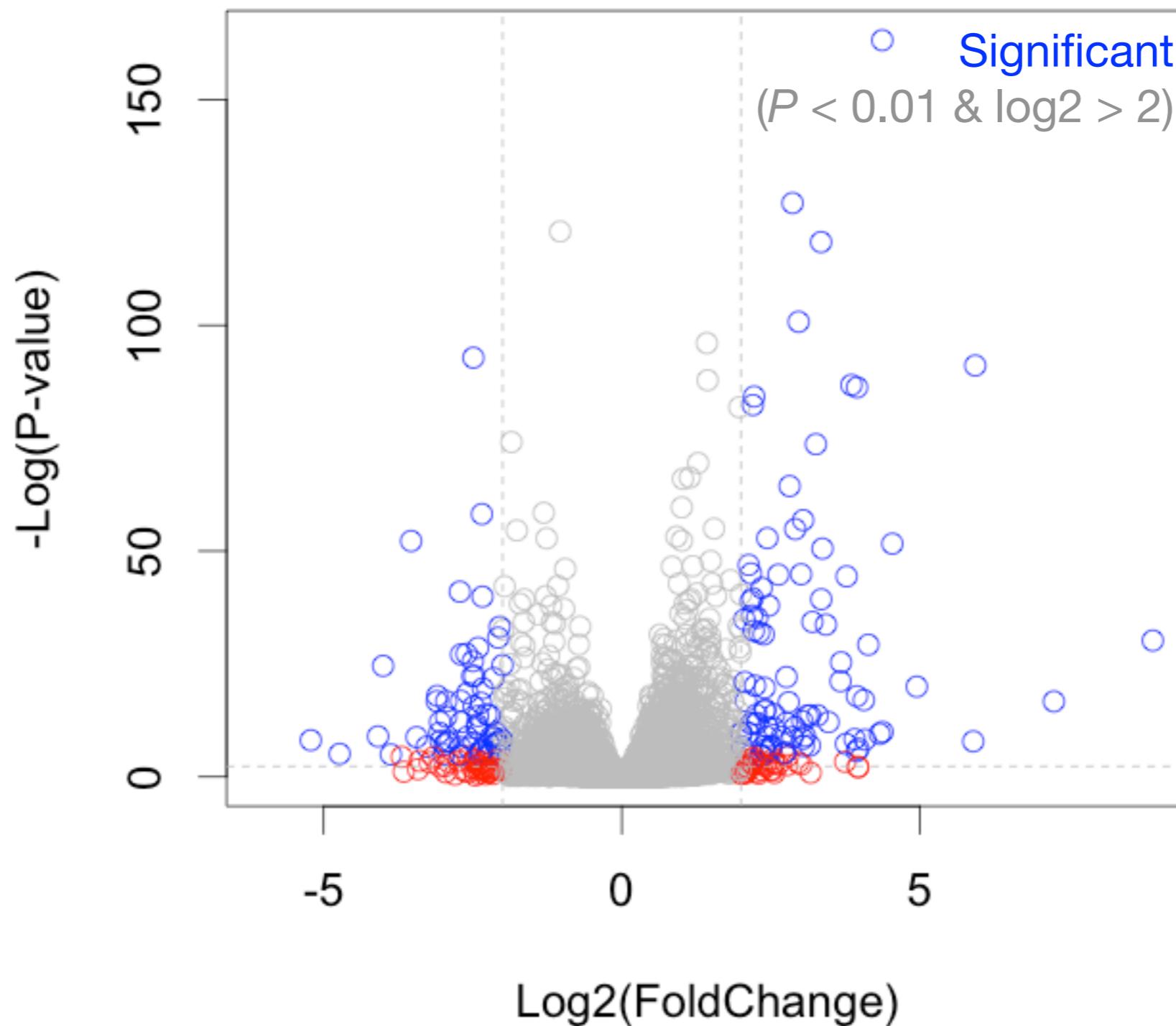
<http://thegrantlab.org/bimm143>



X	baseMean	log2FoldChange	IfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
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

## Volcano Plot

### Fold change vs P-value



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

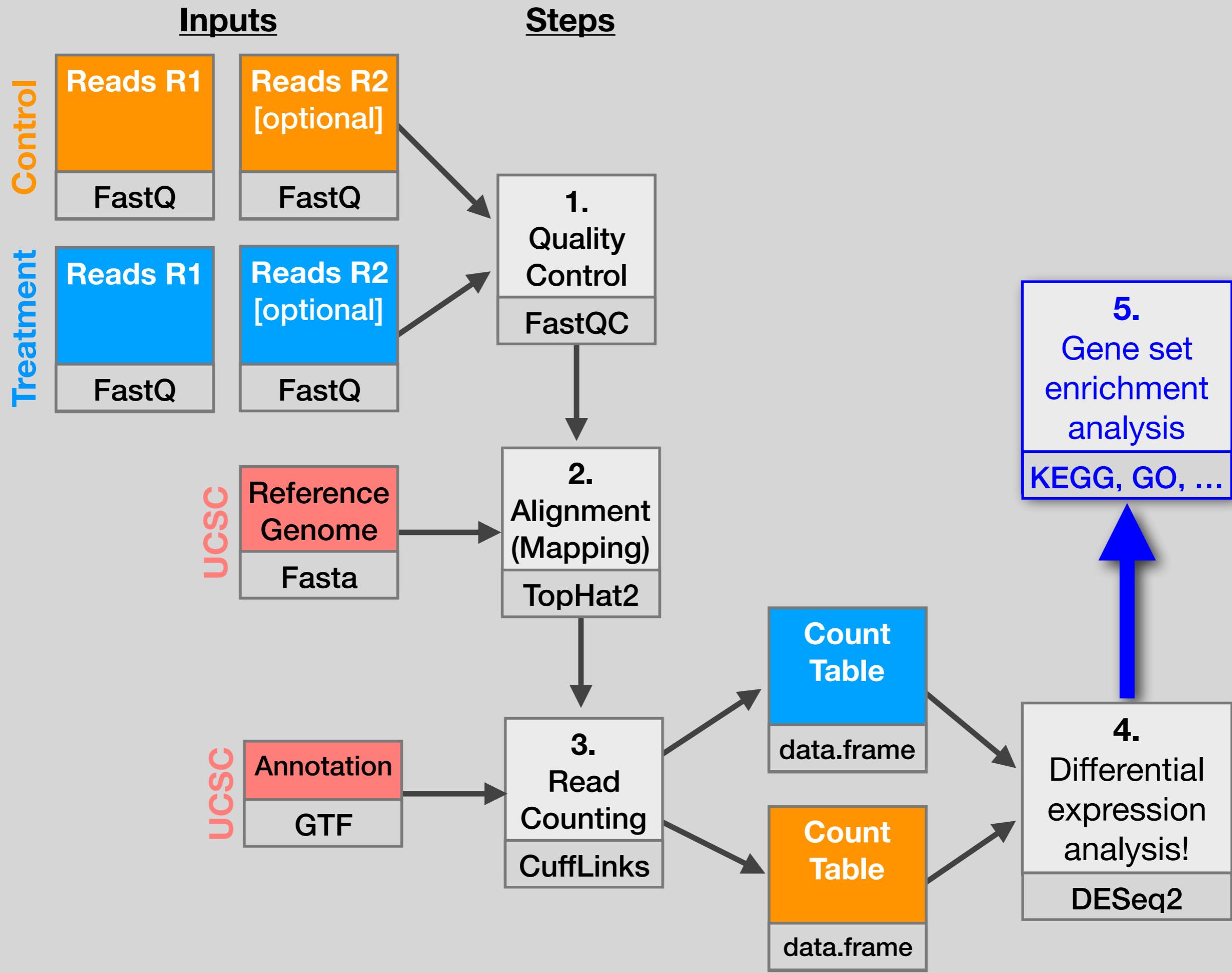
What do I do now?



# Pathway analysis!

(a.k.a. geneset enrichment)

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



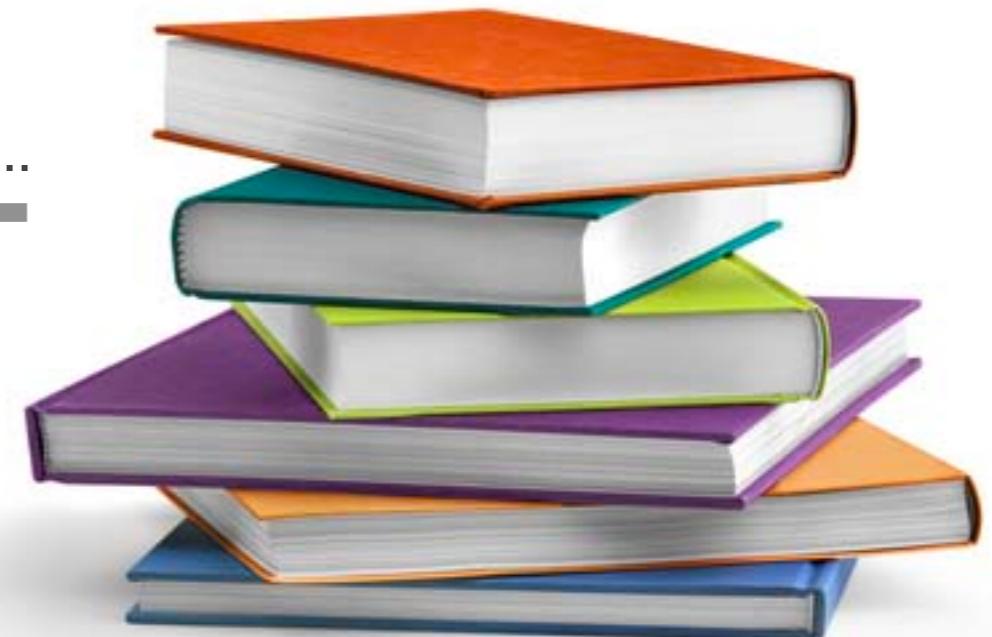
# Basic idea

## Differentially Expressed Genes (DEGs)

X	baseMean	log2FoldChange	IfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
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

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

Annotate...



# Basic idea

## Differentially Expressed Genes (DEGs)

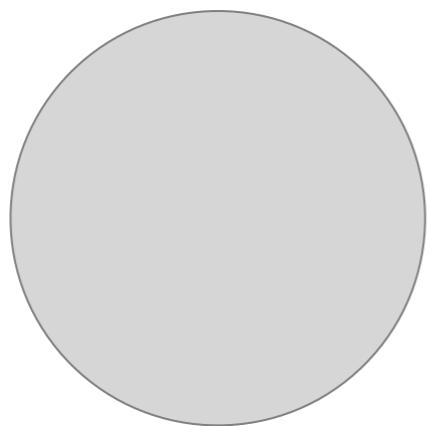
X	baseMean	log2FoldChange	IfcSE	stat	pvalue	padj	symbol
ENSG00000152583	954.77093	4.3683590	0.23713648	18.421286	8.867079e-76	1.342919e-71	SPARCL1
ENSG00000179094	743.25269	2.8638885	0.17555825	16.313039	7.972621e-60	6.037267e-56	PER1
ENSG00000116584	2277.91345	-1.0347000	0.06505273	-15.905557	5.798513e-57	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

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

Annotate...

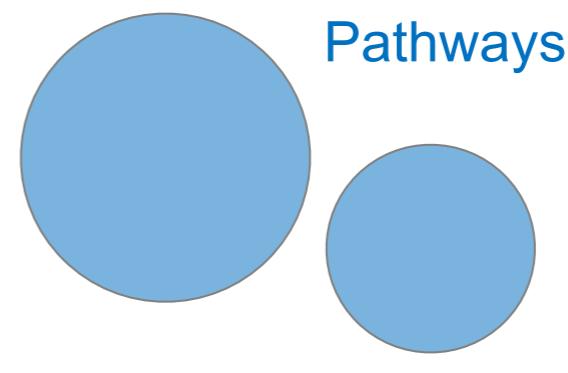


Differentially  
Expressed  
Genes  
(DEGs)



Pathway analysis  
(geneset enrichment)

Overlap...

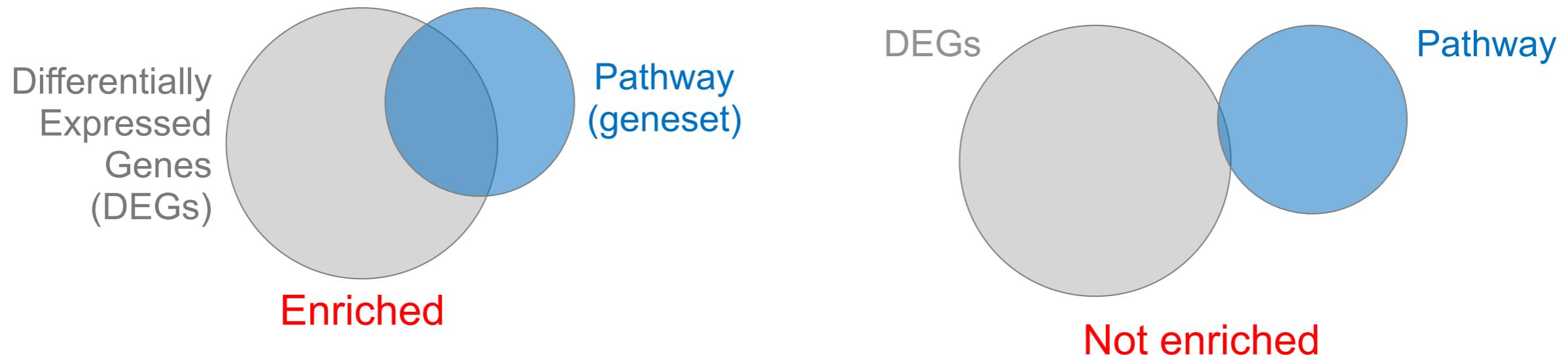


Pathways

# 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*

# 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-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

228018	_at	226	ENSG00000090339	NP_000192	C20orf58
226		207	ENSG00000010030	NP_057219	
		225	ENSG00000110030	055029	
		221	ENSG00000210030	000585	
		1553	ENSG00000210030	006125	
		2184	ENSG00000210030	589495	
		2049	ENSG00000210030	01032249	
		2026	ENSG00000210030	78870	
		23095	ENSG00000210030	4515	
		22801	ENSG00000210030	3839	
		15540	ENSG00000210030	1412	
		20312	ENSG00000210030	069	NMB
		225182	ENSG00000210030	183	PA2
		225079	ENSG00000210030	01	MEM50B
		243010	ENSG00000210030	05340	MP2
		230668	ENSG00000210030	5	MSI2
		218541	ENSG00000210030	4050	C20orf58
		224225	ENSG00000210030	NP_033666	C8orf4
		207339	ENSG00000210030	NP_002332	ETV7
		202637	s_at	W03F8.6	LTB
					ICAM1

# 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 with several key features highlighted:

- Search Bar:** "Search in" dropdown set to "Protein Knowledgebase (UniProtKB)" and a "Query" input field.
- Action Buttons:** "Search", "Clear", "Fields", "Blast", "Align", "Retrieve", and "ID Mapping". The "ID Mapping" button is highlighted with a red box.
- Welcome and News:** "WELCOME" and "NEWS" sections with an RSS icon.
- Identifier Translation Form:** A central form for translating identifiers. It includes:
  - Identifiers:** A large input area with a blue border.
  - From:** A dropdown menu set to "EMBL/GenBank/DDBJ".
  - To:** A dropdown menu set to "UniProtKB AC".
  - File Input:** A "Choose File" button with the message "no file selected".
  - Action Buttons:** "Map", "Swap", and "Clear".

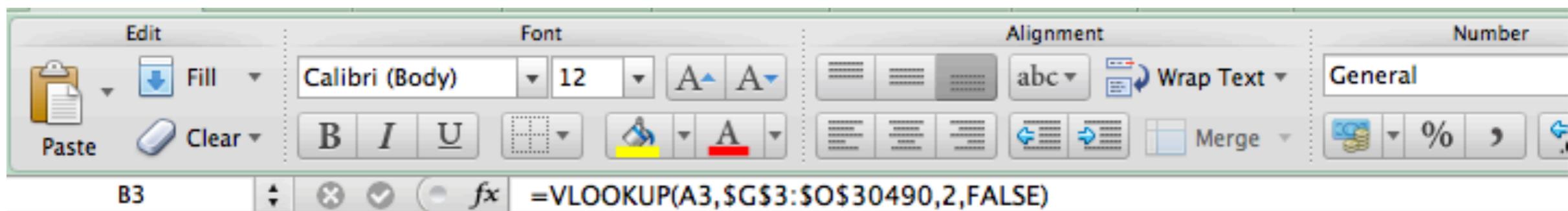
# Translating between identifiers

---

- 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: 'Data Table' and 'Annotation Table'. The 'Data Table' (A1:A17) contains RefSeq, Symbol, and three expression values (Exp1, Exp2, Exp3). The 'Annotation Table' (G1:K17) contains RefSeq, Symbol, Entrez ID, Unigene, and RefSeq. Cell B3 contains the formula =VLOOKUP(A3,\$G\$3:\$O\$30490,2, FALSE), which looks up the RefSeq from the Data Table in the Annotation Table's RefSeq column.

	A	B	C	D	E	F	G	H	I	J	K
1	Data Table						Annotation Table				
2	RefSeq	Symbol	Exp1	Exp2	Exp3		RefSeq	Symbol	Entrez ID	Unigene	RefSeq
3	NM_153103	Kif1c	2.31975457	1.24558927	2.78816871		NM_001001	Zfp85-rs1	22746	Mm.288396	NM_001
4	NM_146017	Gabrp	4.15029735	3.08055836	1.18919962		NM_001001	Scap	235623	Mm.288741	NM_001
5	NM_018883	Camkk1	3.83282512	0.0522951	0.64684259		NM_001001	Scap	235623	Mm.288741	NM_001
6	NM_145936	Tspy12	0.45449369	1.62761318	7.59770627		NM_001001	Fbxo41	330369	Mm.38777	NM_001
7	NM_026599	Cgnl1	4.84541871	2.84751796	1.61595768		NM_001001	Taf9b	407786	Mm.19440	NM_001
8	NM_013926	Cbx8	1.22903318	0.2863077	0.02952665		NM_001001	Taf9b	407786	Mm.19440	NM_001
9	NR_015566	A330023F24	1.44695053	0.98809479	1.59330144		NM_001001	BC051142	407788	Mm.73205	NM_001
10	NM_008623	Mpz	0.50749263	0.94350028	6.10581569		NM_001001	BC051142	407788	Mm.73205	NM_001
11	NM_183127	Fate1	2.45672795	4.87960794	3.60759511		NM_001001	BC048546	232400	Mm.259234	NM_001
12	NM_008943		4.78701069	4.15302647	0.85432314		NM_001001	Zfp941	407812	Mm.359154	NM_001
13	NM_025382		0.66397344	1.40664187	3.09539802		NM_001001	BC031181	407819	Mm.29866	NM_001
14	NM_182841		1.25528938	0.20505996	2.76879488		NM_001001	Baz2b	407823	Mm.486364	NM_001
15	NM_030061		0.17670108	2.75415469	2.98900691		NM_001001	Tmem204	407831	Mm.34379	NM_001
16	NM_133216		6.572343	0.59671282	3.84650536		NM_001001	Ccdc111	408022	Mm.217385	NM_001
17	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](#)]



## # Using the merge() function

&gt; anno &lt;- read.csv("data/annotables\_grch38.csv")

This is an annotation file

&gt; merge(mygenes, anno, by.x="row.names", by.y= "ensgene")

This is our differential expressed genes





## # Using the merge() function

&gt; anno &lt;- read.csv("data/annotables\_grch38.csv")

&gt; merge(mygenes, anno, by.x="row.names", by.y= "ensgene")

## # Using mapIds() function from bioconductor

&gt; library("AnnotationDbi")

&gt; 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", "PDXK",
      "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
```

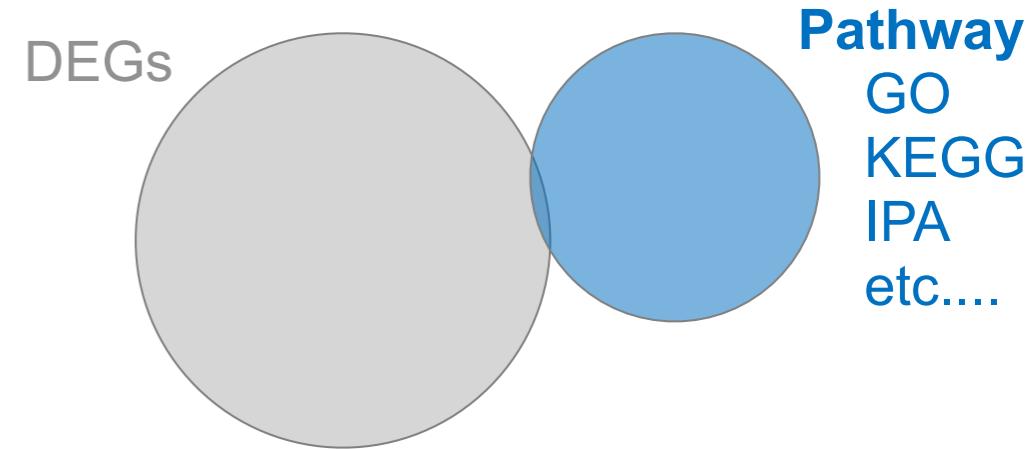
See package vignette:

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

# 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 Gene Ontology

AmiGO

Search   Browse   BLAST   Homolog Annotations   Tools & Resources   Help

Search the Gene Ontology database

GO terms    genes or proteins    exact match

Beta

AmiGO 2

AmiGO version: 1.8

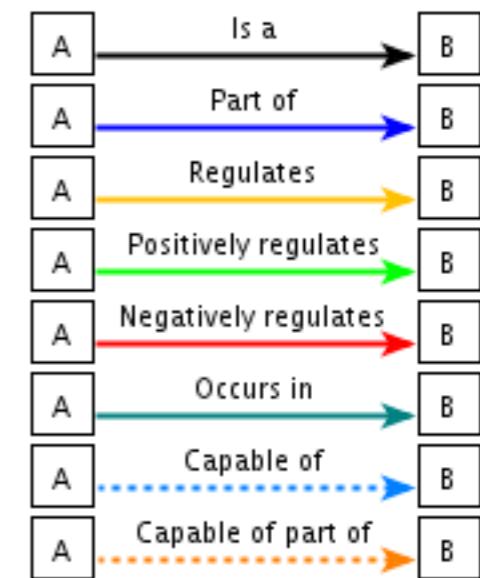
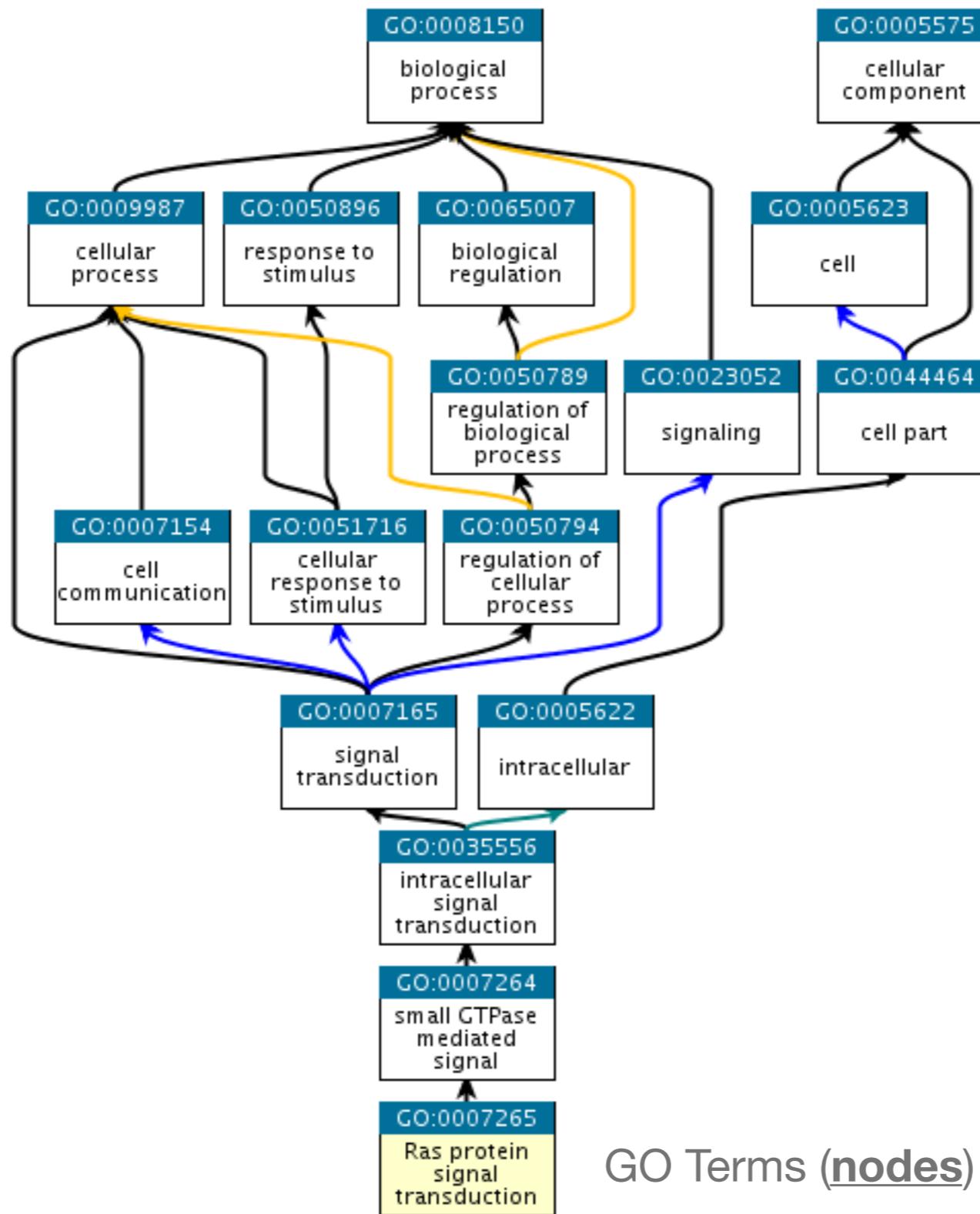
Try AmiGO Labs

GO database release 2013-10-05

Cite this data • Terms of use • GO helpdesk

Copyright © 1999-2010 the Gene Ontology

# GO is structured as a “directed graph”



Relationships (edges)

Parent terms are more general & child terms more specific

# GO evidence codes

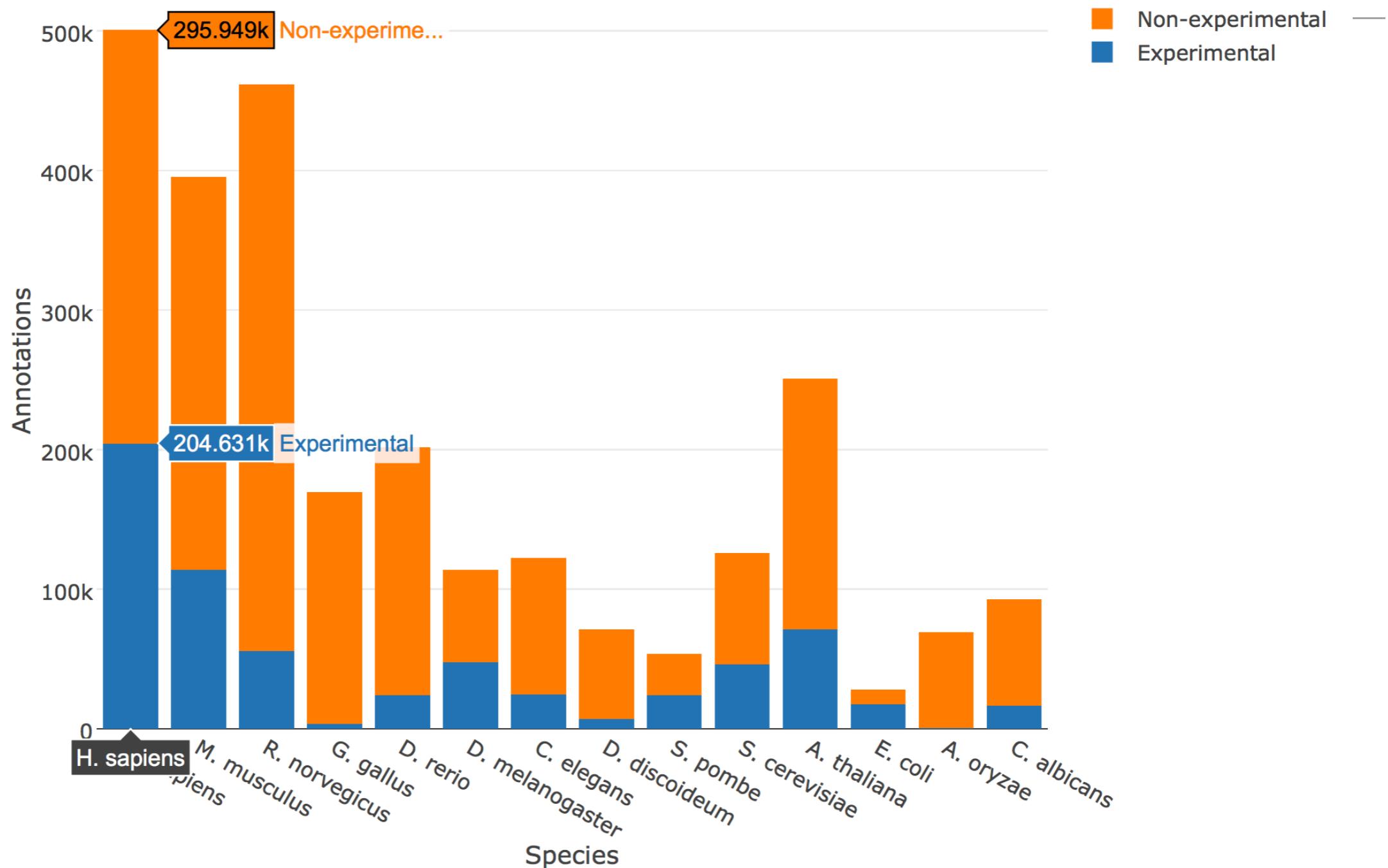
Evidence code	Evidence code description	Source of evidence	Manually checked	Current number of annotations*
IDA	Inferred from direct assay	Experimental	Yes	71,050
IEP	Inferred from expression pattern	Experimental	Yes	4,598
IGI	Inferred from genetic interaction	Experimental	Yes	8,311
IMP	Inferred from 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

## Use and misuse of the gene ontology annotations

Seung Yon Rhee, Valerie Wood, Kara Dolinski & Sorin Draghici  
*Nature Reviews Genetics* 9, 509-515 (2008)

## Experimental annotations by species



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

# Can now do gene list analysis with GeneGO online!

Screenshot of the PANTHER Classification System website showing the Gene List Analysis tool.

The URL in the browser is [pantherdb.org/webservices/go/overrep.jsp](http://pantherdb.org/webservices/go/overrep.jsp).

The page features the PANTHER Classification System logo (a black panther silhouette) and the GENEONTOLOGY Unifying Biology logo.

The navigation menu includes: Home, About, PANTHER Data, PANTHER Tools, Workspace, Downloads, Help/Tutorial, LOGIN, REGISTER, and CONTACT US.

A banner at the top states: "New! PANTHER13.1 released."

The main content area has tabs for Gene List Analysis, Browse, Sequence Search, cSNP Scoring, and Keyword Search. The Gene List Analysis tab is active.

A message in the center says: "Please refer to our article in [Nature Protocols](#) for detailed instructions on how to use this page." Below it, an error message reads: "Error parsing request, no input specified".

The "Gene List Analysis" section contains three numbered steps:

- Enter IDs:** A text input field for "Supported IDs" with a note: "separate IDs by a space or comma". Below it is a "Choose File" button with "no file selected".
- Select List Type:** Radio buttons for:
  - ID List (selected)
  - Previously exported text search results
  - Workspace list
  - PANTHER Generic Mapping File
  - VCF FileA dropdown menu for "Flanking region" is set to "20 Kb".
- Select organism:** A dropdown menu showing options: Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, and Danio rerio.

The "Gene List Analysis" section also includes a "Help Tips" box with steps:

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

On the left sidebar, under "Quick links", are: Whole genome function views, Genome statistics, Data Version, How to cite PANTHER, and NEW! Recent publication describing PANTHER.

Under "News", it says PANTHER13.1 Released and there is a link to Click for additional info.

Under "Newsletter subscription", there is a form to enter an email address and a "Subscribe" button. A PostgreSQL logo is also present.

# Another popular online tool: **DAVID** at NIAID <[david.abcc.ncifcrf.gov](http://david.abcc.ncifcrf.gov)>

**DAVID Bioinformatics Database**

**Analysis Wizard**  
DAVID Bioinformatics Resources 2008, NIAID/NIH

[Home](#) [Start Analysis](#) [Shortcut to DAVID Tools](#) [Technical Center](#) [Downloads & APIs](#) [Term of Service](#) [Why DAVID?](#) [About Us](#)

**Upload Gene List**

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

[Upload Help](#)

**Step 1: Enter Gene List**

A: Paste a list

Box A (empty) [Clear](#)

Or

B: Choose From a File

[Choose File](#) no file selected

**Step 2: Select Identifier**

[AFFY\\_ID](#)

**Step 3: List Type**

Gene List  Background

**Step 4: Submit List**

[Submit List](#)

**Analysis Wizard**

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

Step 1. Submit your gene list through left panel.

<sup>new!</sup>Note: Affy Exon IDs and Affy Gene Array IDs are now supported in DAVID, as "affy\_id" type.

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

# DAVID

- *Functional Annotation Chart*

**Functional Annotation Chart**

Current Gene List: Uploaded List\_1  
Current Background: Homo sapiens  
2316 DAVID IDs

Help and Manual

Options

Rerun Using Options Create Sublist

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">regulation of progression through cell cycle</a>	<a href="#">RT</a>		98	4.2	3.3E-7	8.6E-4
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">apoptosis</a>	<a href="#">RT</a>		131	5.7	1.6E-6	2.1E-3
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">cell death</a>	<a href="#">RT</a>		136	5.9	3.8E-6	3.3E-3
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">regulation of transcription from RNA polymerase II promoter</a>	<a href="#">RT</a>		83	3.6	3.7E-5	2.4E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">protein kinase cascade</a>	<a href="#">RT</a>		71	3.1	4.7E-5	2.4E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">regulation of kinase activity</a>	<a href="#">RT</a>		48	2.1	5.4E-5	2.3E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">negative regulation of cell proliferation</a>	<a href="#">RT</a>		48	2.1	1.0E-4	3.7E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">regulation of cell size</a>	<a href="#">RT</a>		41	1.8	1.2E-4	3.9E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">monocarboxylic acid metabolic process</a>	<a href="#">RT</a>		48	2.1	1.3E-4	3.6E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process</a>	<a href="#">RT</a>		61	2.6	1.5E-4	3.8E-2
<input type="checkbox"/>	GOTERM_BP_5	<a href="#">positive regulation of cellular metabolic process</a>	<a href="#">RT</a>		72	3.1	1.7E-4	3.8E-2

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:

## Annotation Summary Results

[Help and Tool Manual](#)

**Current Gene List:** Uploaded List\_3  
**Current Background:** HOMO SAPIENS

**Main Accessions** (0 selected)  
 **Other Accessions** (0 selected)  
 **Gene Ontology** (4 selected)  
 **Protein Domains** (3 selected)  
 **Pathways** (3 selected)  
 **General Annotations** (0 selected)  
 **Functional Categories** (3 selected)  
 **Protein Interactions** (0 selected)  
 **Literature** (0 selected)  
 **Disease** (1 selected)  
 **Tissue Expression**

**2320 DAVID IDs**  
 **Check Defaults**

### Combined View for Selected Annotation



# DAVID Functional Annotation Clustering

- Based on shared genes between functional sets

**Functional Annotation Clustering**

[Help and Manual](#)

**Current Gene List: Uploaded List\_3**  
**2320 DAVID IDs**

**Options**    **Classification Stringency**

Annotation Cluster 1	Enrichment Score: 3.72	G		Count	P_Value	Benjamini
<input type="checkbox"/> GOTERM_BP_5	<a href="#">regulation of transcription from RNA polymerase II promoter</a>	RT		83	3.7E-5	2.4E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process</a>	RT		61	1.5E-4	3.8E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of cellular metabolic process</a>	RT		72	1.7E-4	3.8E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of transcription</a>	RT		58	3.8E-4	5.0E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of transcription, DNA-dependent</a>	RT		48	7.4E-4	7.6E-2
Annotation Cluster 2	Enrichment Score: 3.54	G		Count	P_Value	Benjamini
<input type="checkbox"/> GOTERM_BP_5	<a href="#">regulation of cell size</a>	RT		41	1.2E-4	3.9E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">regulation of cell growth</a>	RT		33	3.7E-4	5.1E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">cell morphogenesis</a>	RT		81	5.2E-4	5.7E-2
Annotation Cluster 3	Enrichment Score: 3.37	G		Count	P_Value	Benjamini
<input type="checkbox"/> GOTERM_BP_5	<a href="#">apoptosis</a>	RT		131	1.6E-6	2.1E-3
<input type="checkbox"/> GOTERM_BP_5	<a href="#">cell death</a>	RT		136	3.8E-6	3.3E-3
<input type="checkbox"/> GOTERM_BP_5	<a href="#">regulation of programmed cell death</a>	RT		88	3.2E-4	5.8E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of apoptosis</a>	RT		48	3.3E-4	5.6E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">regulation of apoptosis</a>	RT		87	3.5E-4	5.2E-2
<input type="checkbox"/> GOTERM_BP_5	<a href="#">positive regulation of programmed cell death</a>	RT		48	4.0E-4	5.0E-2

# Want more?

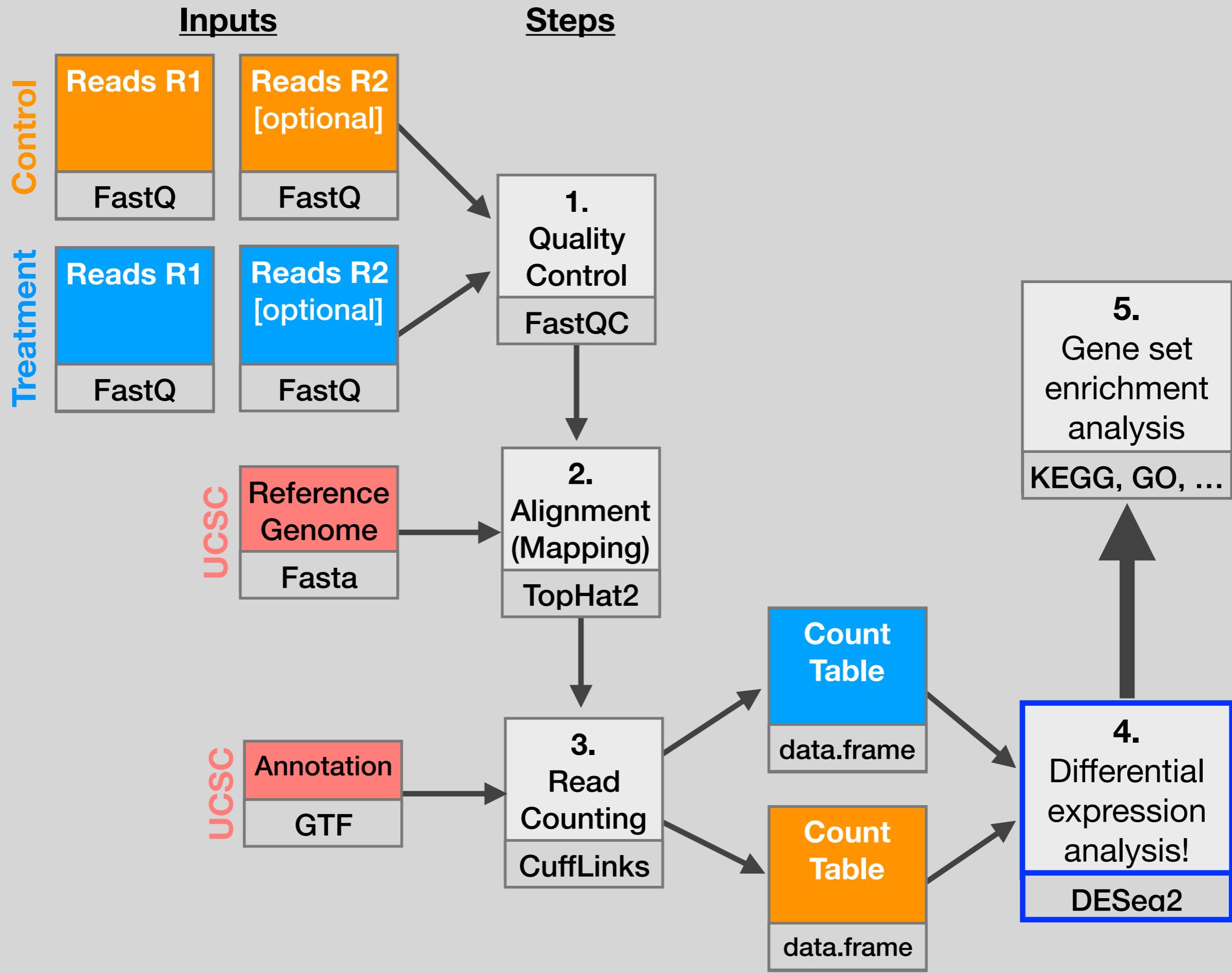
---



- **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!!!**

# Hands-on time!

[https://bioboot.github.io/bimm143\\_F19/lectures/#15](https://bioboot.github.io/bimm143_F19/lectures/#15)



# 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
...	...	...	...	...

2

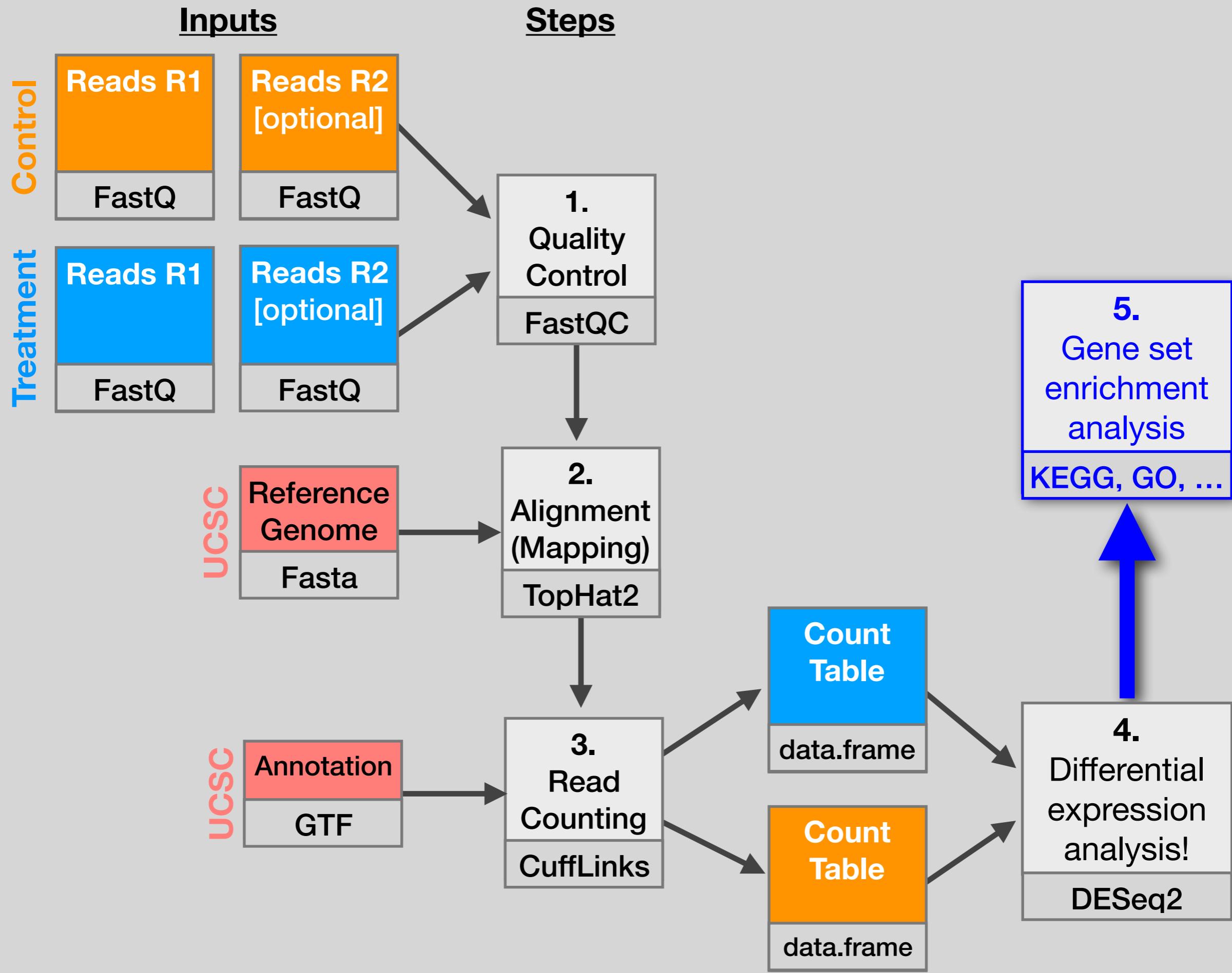
colData

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

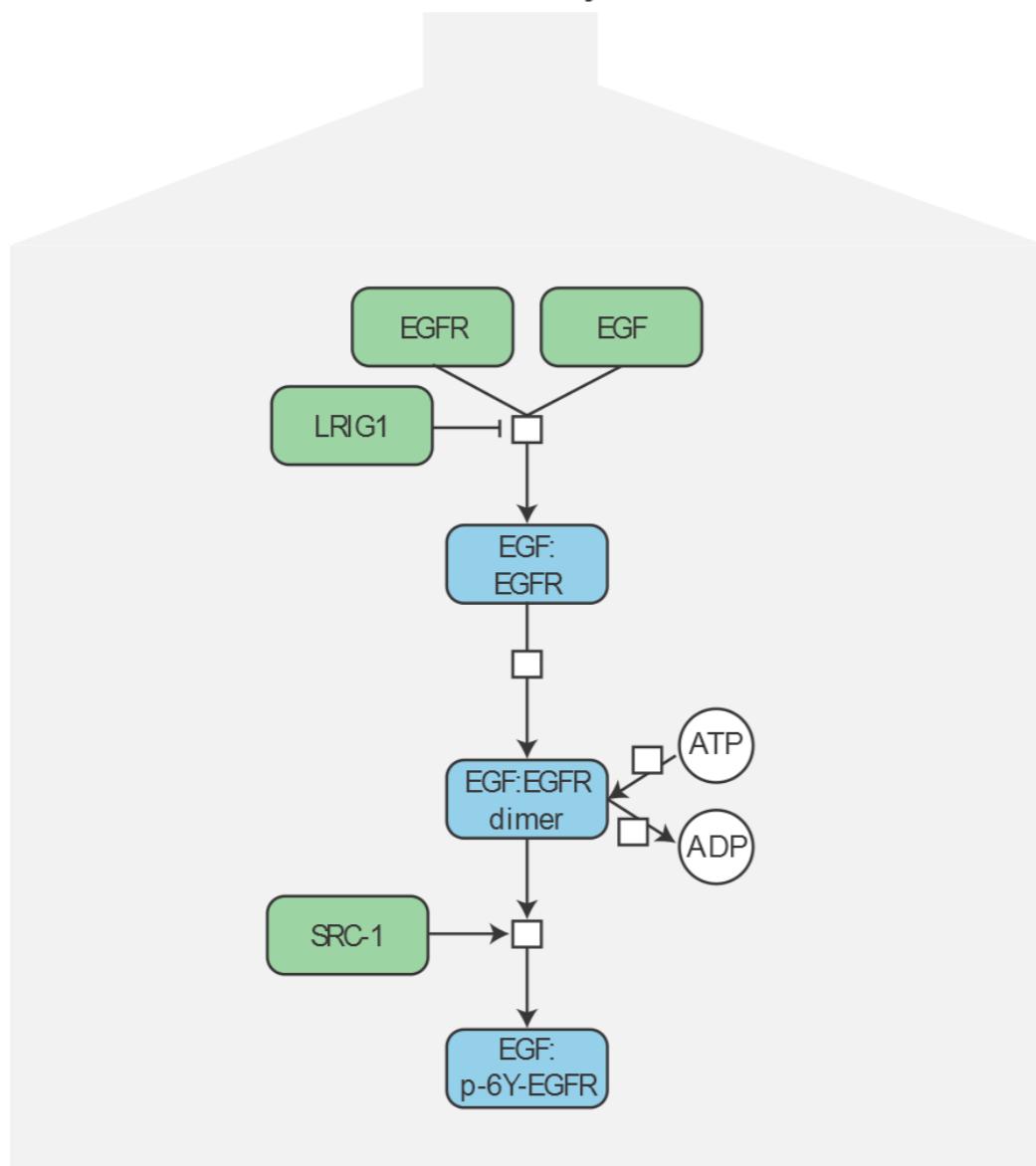
countData is the count matrix  
(Number of reads coming from each  
gene for each sample)

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



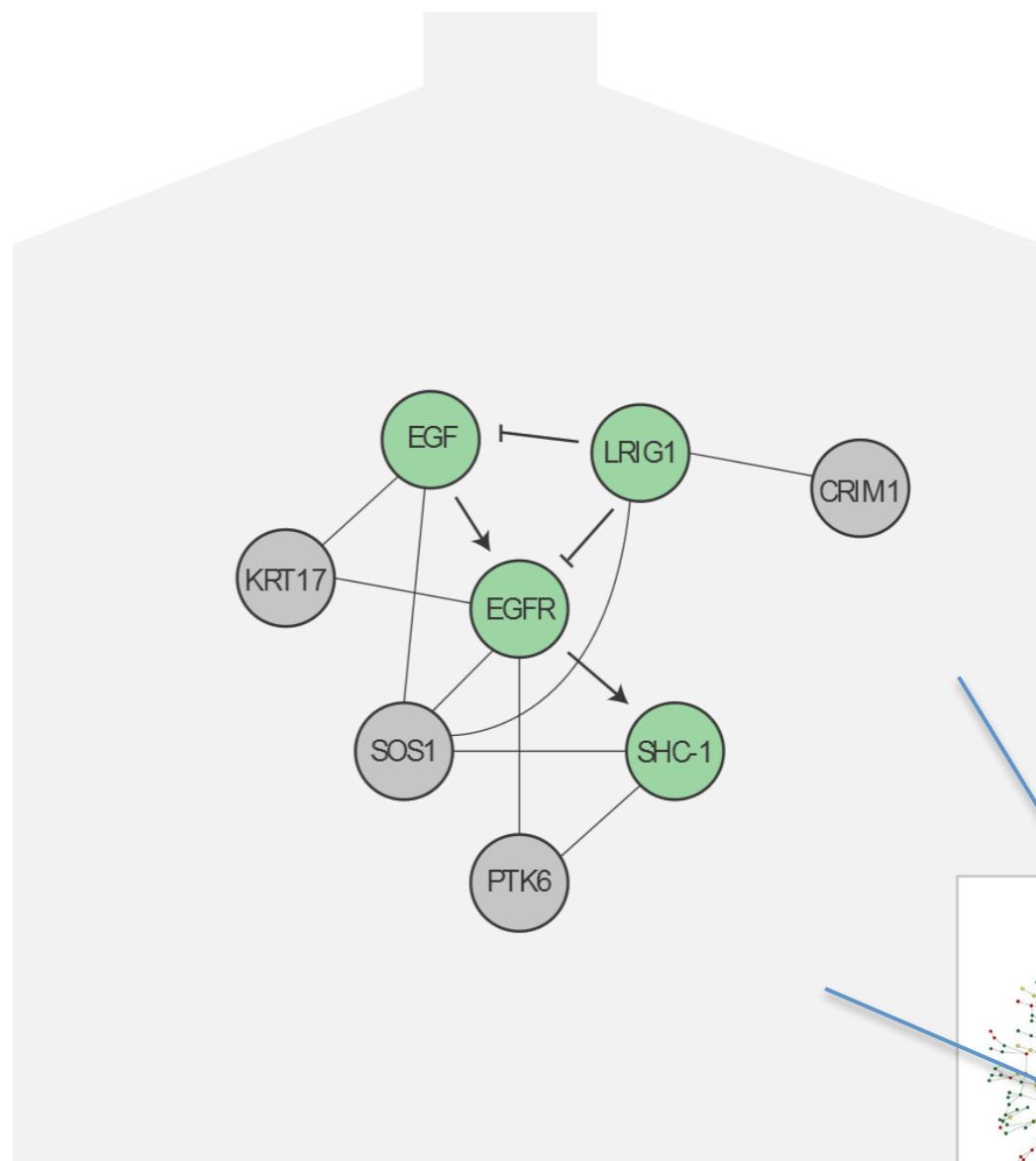
# Pathways vs Networks

EGFR-centered  
Pathway

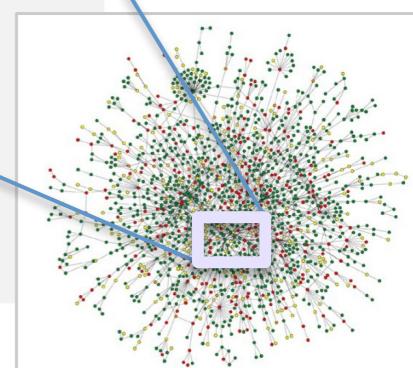


- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

EGFR-centered  
Network



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



# Goal

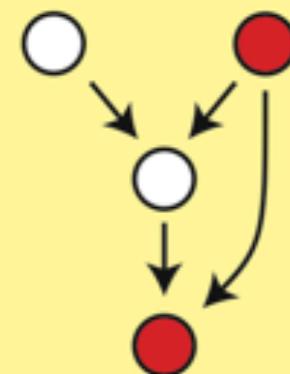
## 1 Enrichment of fixed gene sets

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

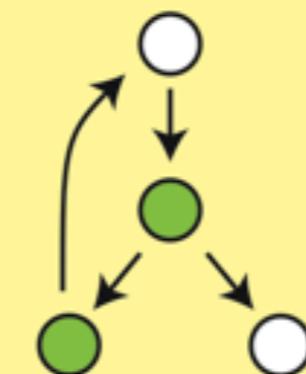
## 2 De novo sub-network construction and clustering

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

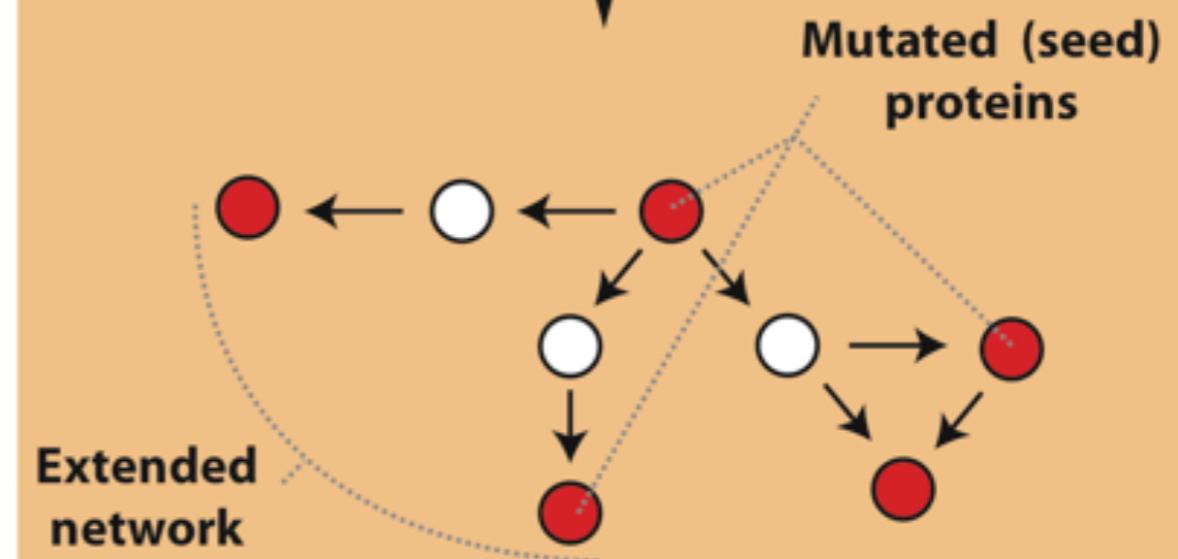
# Output



Enriched network



Depleted network



Extended network

# Goal

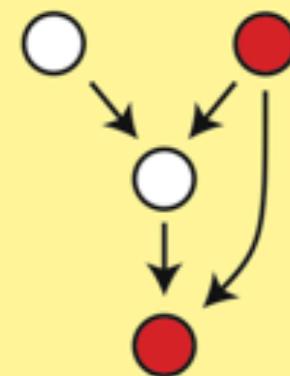
## 1 Enrichment of fixed gene sets

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

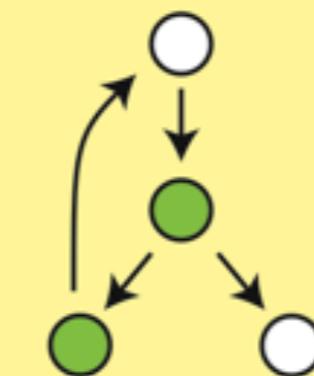
## 2 De novo sub-network construction and clustering

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

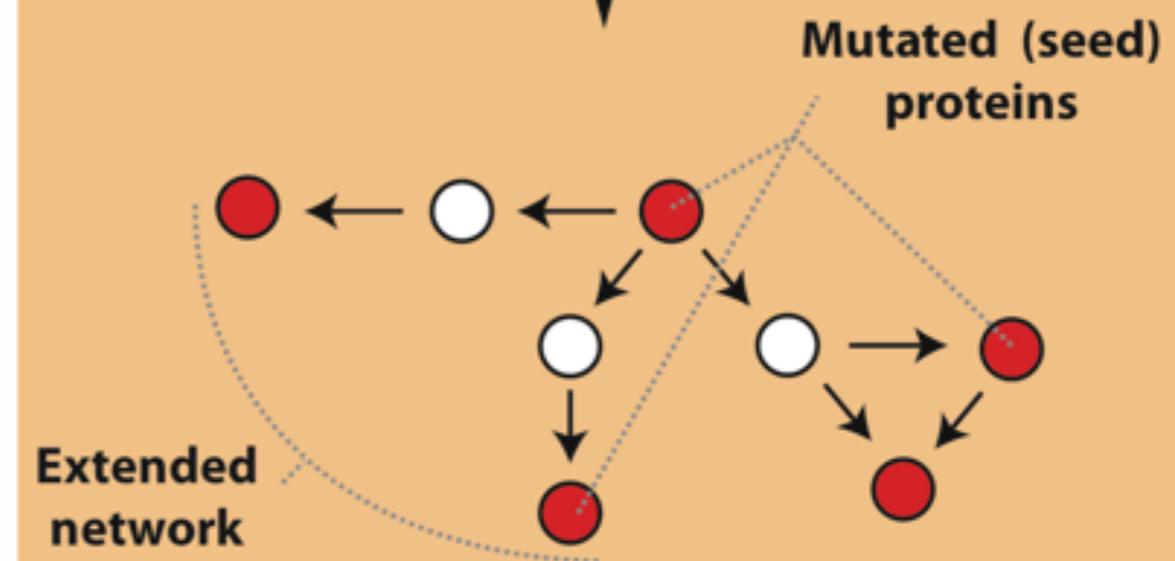
# Output



Enriched network



Depleted network



Extended network

What biological process is altered in this cancer?

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

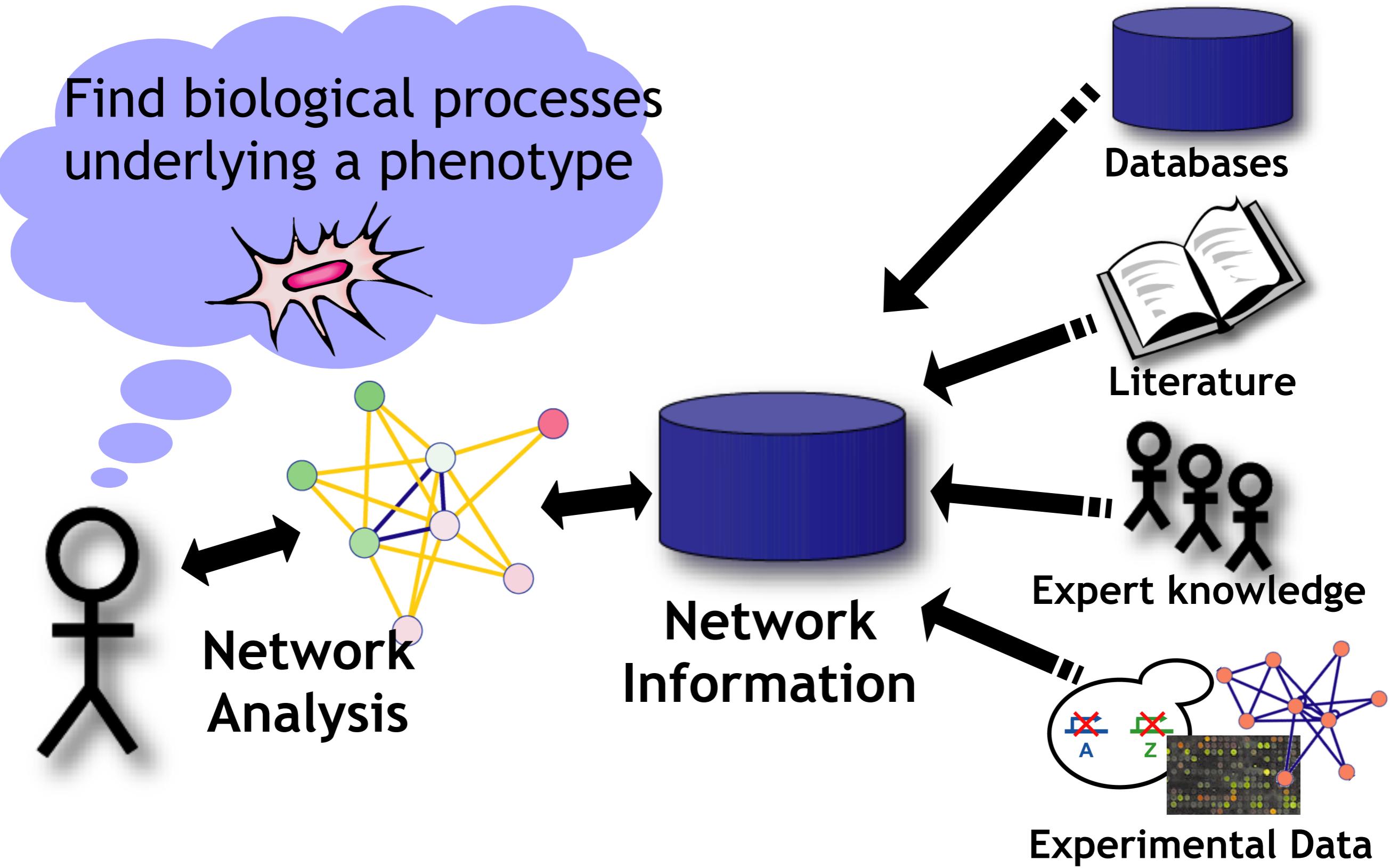
# Pathway analysis (a.k.a. geneset enrichment)

## 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-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

# Pathway & Network Analysis Overview



# R Knowledge Check For BIMM-143

## Quiz

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

Time Limit: 40 mins

