RNA-seq workshop

Annotation data and gene set analysis

Mik Black & Ngoni Faya
Genomics Aotearoa

Annotation

- From wikipedia.org: "Annotation is extra information associat with a particular point in a document or other piece information."
- Here our "document" is the genome.
- The goal of annotating the genome is to link all information relat to sequences, genes, protein, function...

Gene annotation: the good old days...

- Once a fragment of a gene had been sequenced, it was assigne unique identifier called an accession number.
- The accession number was (and still is) used to track that sequen with additional information (e.g., function, other sequenc becoming associated with it as more is learned.
- Sequences used as probes on microarrays can be tracked back an accession number.
- Databases of these identifiers are maintained by the Natio Center for Biotechnology Information (NCBI), and others.

Entrez Gene

- With the sequencing of the human genome, individual sequer fragments could be mapped to their position in the genome, a annotated (in conjunction with other fragments) as "genes".
- Each putative gene in the genome was also assigned an identifie the "Entrez gene" database (also at NCBI).
- The accession numbers of the constituent fragments were th associated with this identifier (and vice versa).
- The gene identifier is also linked to a more descriptive gene nar This usually conveys some information about what that gene dc (or at least what it was understood to be involved in at the tim was named).

Transcriptomics (microarrays and RNA-seq)

- Each spot on a microarray (or transcript fragment in an RNA-s experiment) is associated with a gene (or some sort of meaning DNA sequence), and can thus be linked to an accession numb and/or Entrez ID.
- Through these numbers we can find out about the gene that t seqeunce is associated with (if known), where in the genome it located, and (maybe) what it does.
- In transcriptomic experiments this means that we can find out t identity of genes that undergo differential expression.
- Depending on what is known about these genes, this informati may provide important clues about the underlying biologi process being studied.

Gene function

- For biologists, it's not neccessarily interesting to find out that a ge is significantly differentially expressed if no other information known about that gene.
- One (very good) reason for this is that in transcriptor experiments there are often a lot of false positives, so biologi tend to be a little bit skeptical...
    - Remember: we can only have as much faith in the analysis we do in the underlying assumptions. Were those ger REALLY independent? How about the residuals - norma distributed?

Gene function

- If enough is known about a differentially expressed gene for it "make sense" or be "interesting" in the context of the experime then biologists tend to get a bit more excited.
- Although a gene name is often somewhat informative, v amounts of information about that gene may reside in jour publications and internet databases - how do we get t information?

PubMed identifiers

- PubMed is a service provided by the National Library of Medicine
- Contains over 30 million citations from MEDLINE and other sciences publications.
- Every journal publication is given a unique PubMed identifier.
- Those that relate to a particular gene or sequence are linked ba to the appropriate identifiers.
- Based on this the NCBI search engine hosts a local copy of the NC databases) can be used to retrieve information about differentia expressed genes.

Problem - too much information

- For situations where large numbers of genes are differentia expressed, there is simply too much information available.
- Anyway, are we really interested in individual genes?
- Wouldn't it be better to find groups of differentially express genes which share a common function?

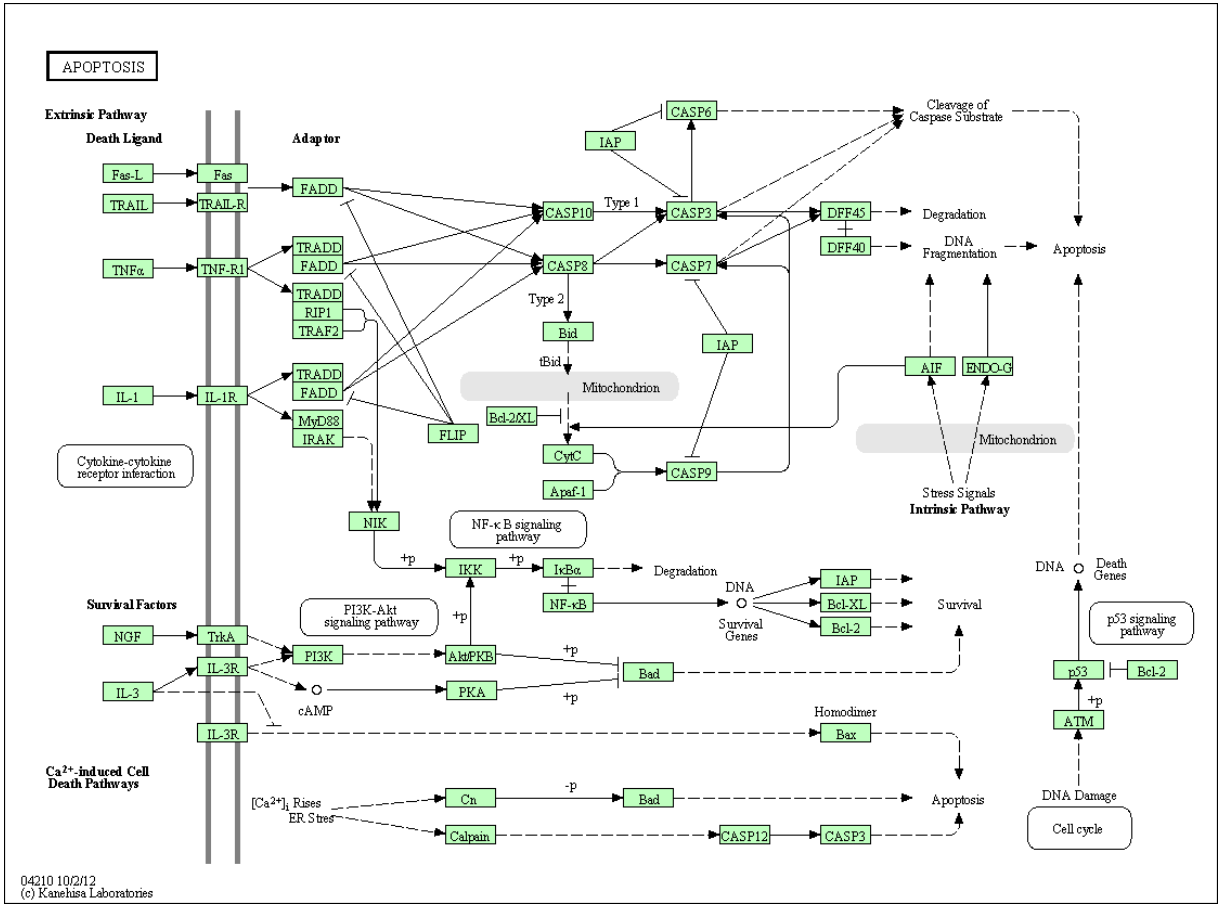
Biological pathways

- In reality genes are members of pathways, which perform ma biological functions.
- As more biological experimentation is done, researchers are able build a better picture of how genes interact, and how pathwa function.
- Information about pathway membership and gene function a stored in publicly available databases.
- This information can be used to define gene sets (groups of ger which are functionally related), to which statistical analysis can applied.

Biological pathways: KEGG

- Kyoto Encyclopedia of Gene and Genomes: http://www.genome.jp/kegg/kegg4.html
- Provides nice (user-created) pathway diagrams (although this is old database, so the style of the diagrams looks a bit dated).
- XML output includes information about genes involved in pathwa and inter-gene (and gene product) relationships.
    - Can produce graphic representation of pathway based on X alone.

KEGG pathway diagram (apoptosis)



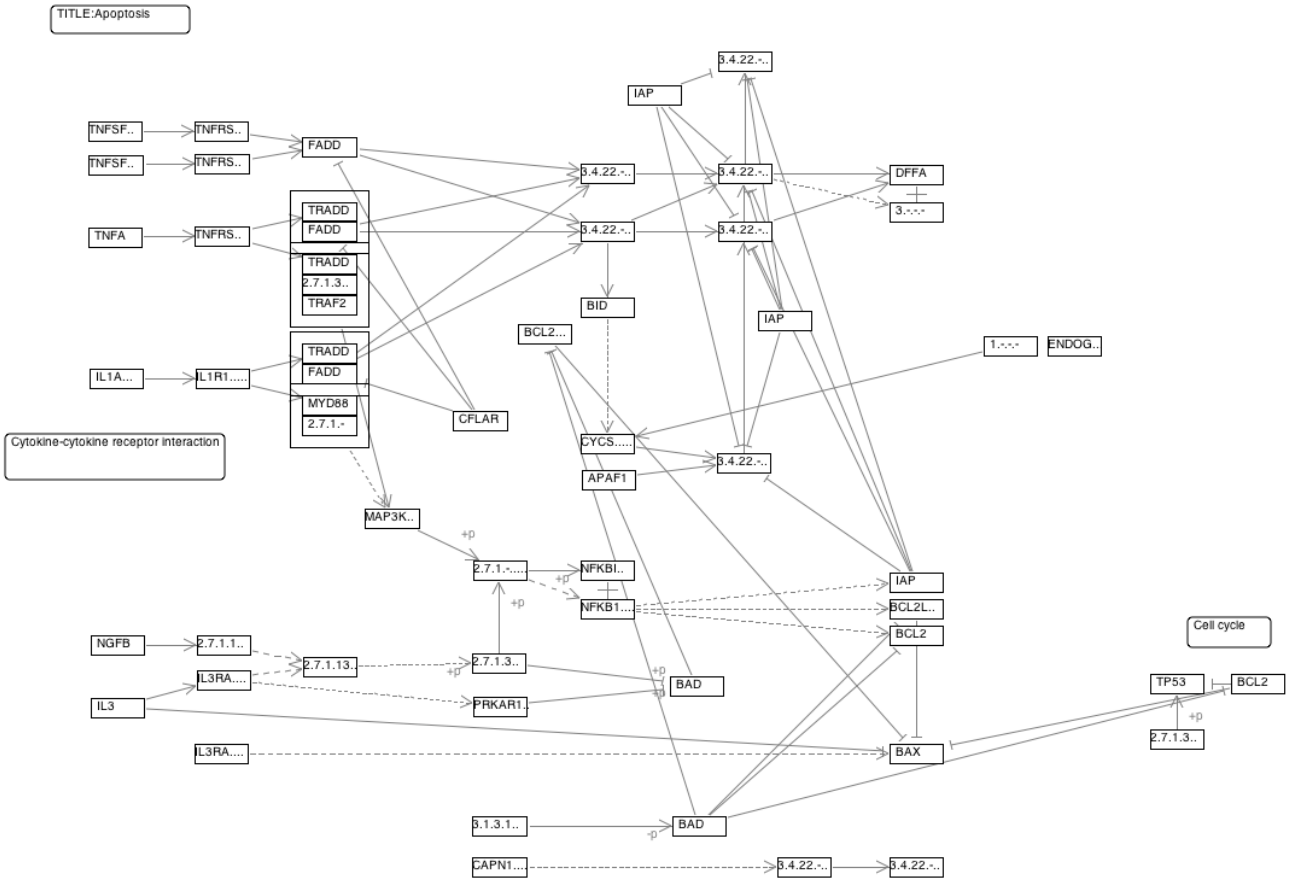
http://www.genome.jp/kegg-bin/show_pathway?org_name=hsa&mapno=04210&mapscale=&

XML output file for apoptosis pathway

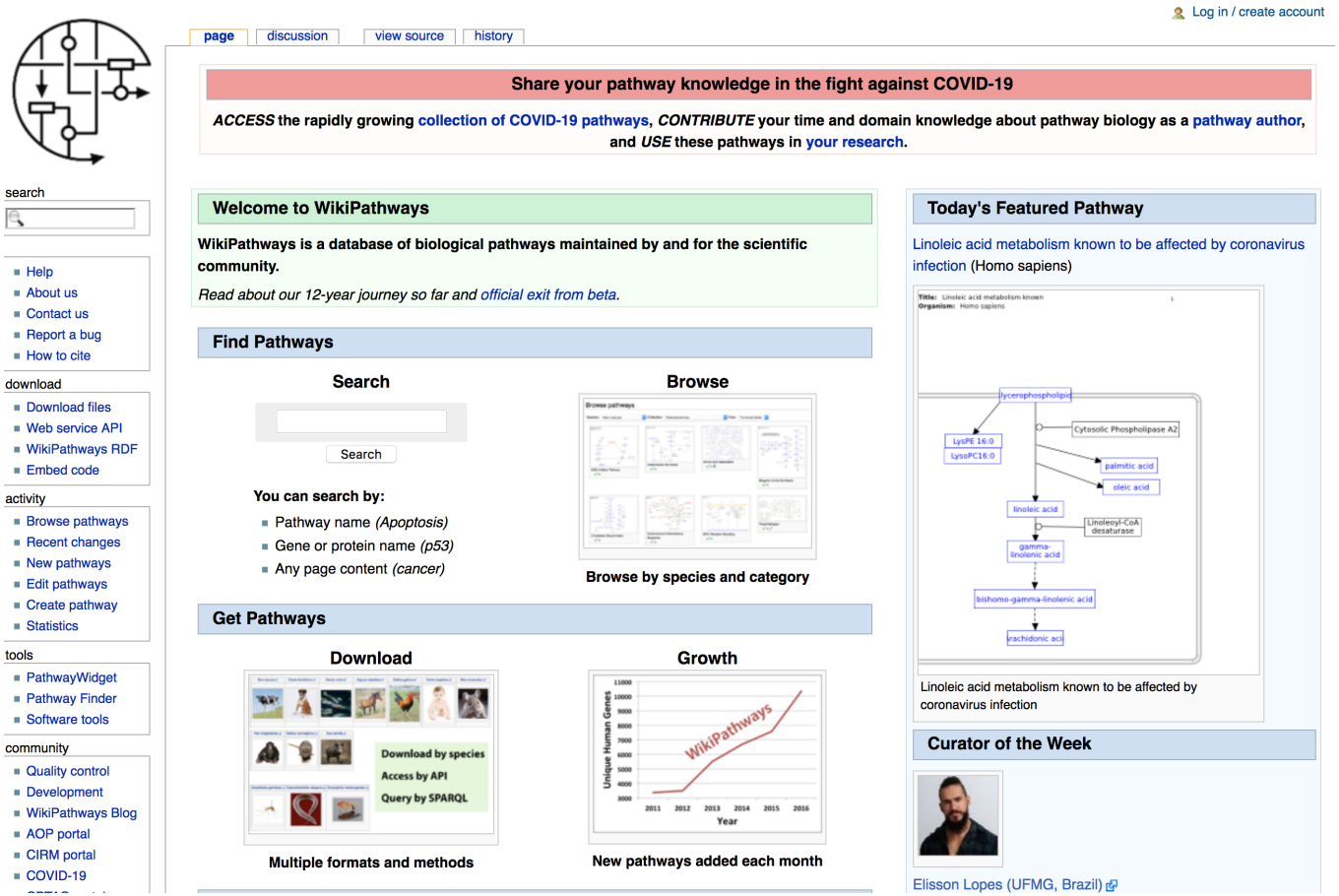
```
:<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd">
<!-- Creation date: Oct 2, 2012 11:48:00 +0900 (GMT+09:00) -->
<pathway name="path:hsa04210" org="hsa" number="04210"
         title="Apoptosis"
         image="http://www.kegg.jp/kegg/pathway/hsa/hsa04210.png"
         link="http://www.kegg.jp/kegg-bin/show_pathway?hsa04210">
    <entry id="1" name="path:hsa04115" type="map"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa04115">
        <graphics name="p53 signaling pathway" fgcolor="#000000" bgcolor="#FFFFFF"
            type="roundrectangle" x="1049" y="572" width="95" height="39"/>
    </entry>
    <entry id="2" name="path:hsa04060" type="map"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa04060">
        <graphics name="Cytokine-cytokine receptor interaction" fgcolor="#000000" bgcolor="#FFFFFF"
            type="roundrectangle" x="111" y="427" width="124" height="39"/>
    </entry>
    <entry id="3" name="hsa:5530 hsa:5532 hsa:5533 hsa:5534 hsa:5535" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:5530+hsa:5532+hsa:5533+hsa:5534+hsa:5535">
        <graphics name="PPP3CA, CALN, CALNA, CALNA1, CCN1, CNA1, PPP2B..." fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="430" y="733" width="46" height="17"/>
    </entry>
    <entry id="4" name="hsa:581" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:581">
        <graphics name="BAX, BCL2L4" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="776" y="673" width="46" height="17"/>
    </entry>
    <entry id="5" name="hsa:598" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:598">
        <graphics name="BCL2L1, BCL-XL/S, BCL2L, BCLX, BCLXL, BCLXS, Bcl-X, PPP1R52, bcl-xL, bcl-xS" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="776" y="553" width="46" height="17"/>
    </entry>
    <entry id="6" name="hsa:1676" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:1676">
        <graphics name="DFFA, DFF-45, DFF1, ICAD" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="776" y="192" width="46" height="17"/>
    </entry>
    <entry id="7" name="hsa:596" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:596">
        <graphics name="BCL2, Bcl-2, PPP1R50" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="1059" y="615" width="46" height="17"/>
    </entry>
    <entry id="8" name="hsa:472" type="gene"
        link="http://www.kegg.jp/dbget-bin/www_bget?hsa:472">
```

http://www.kegg.jp/kegg-bin/download?entry=hsa04210&format=kgml

XML-based KEGG diagram (apoptosis)

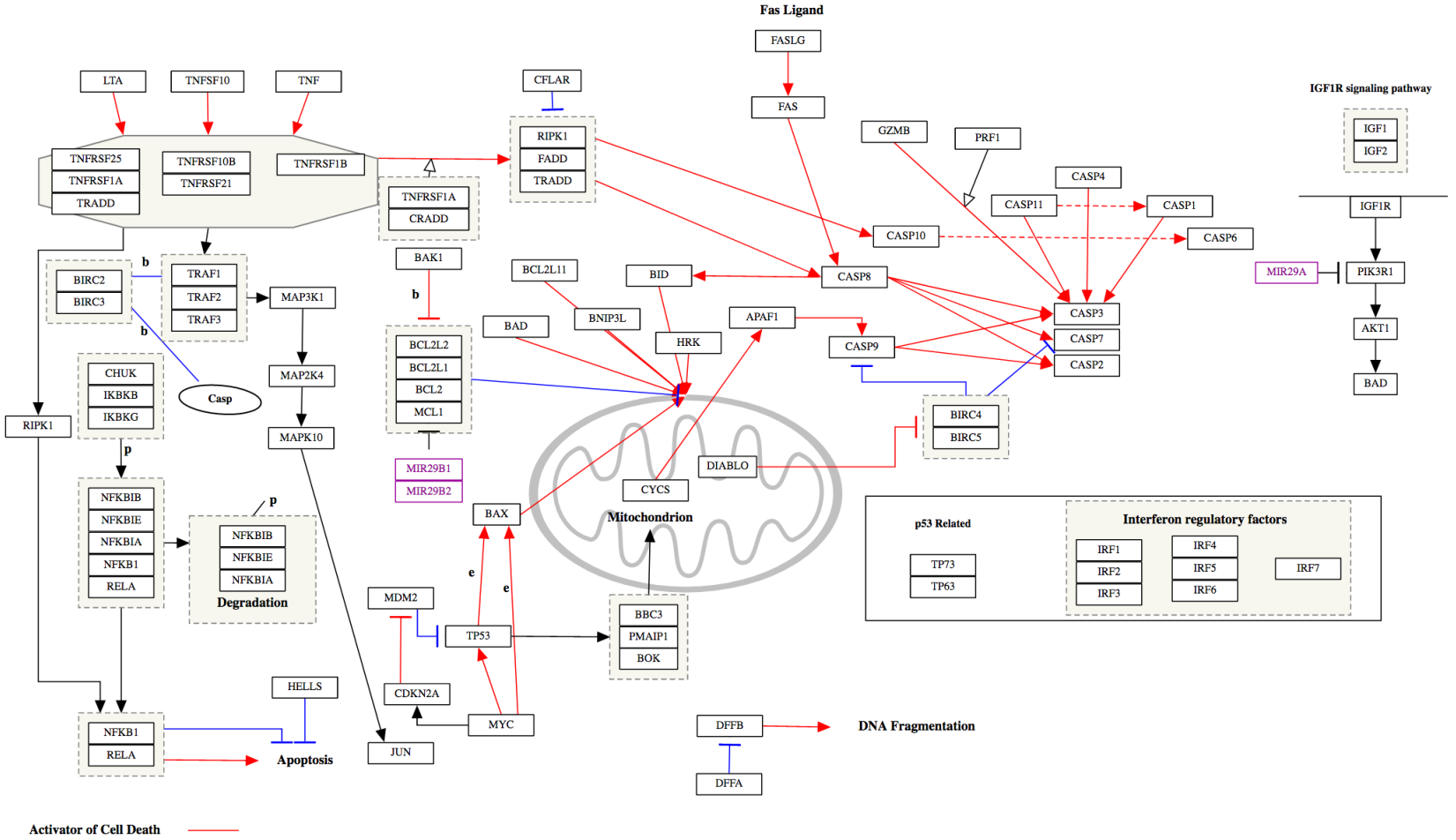


User-curated database: Wikipathways



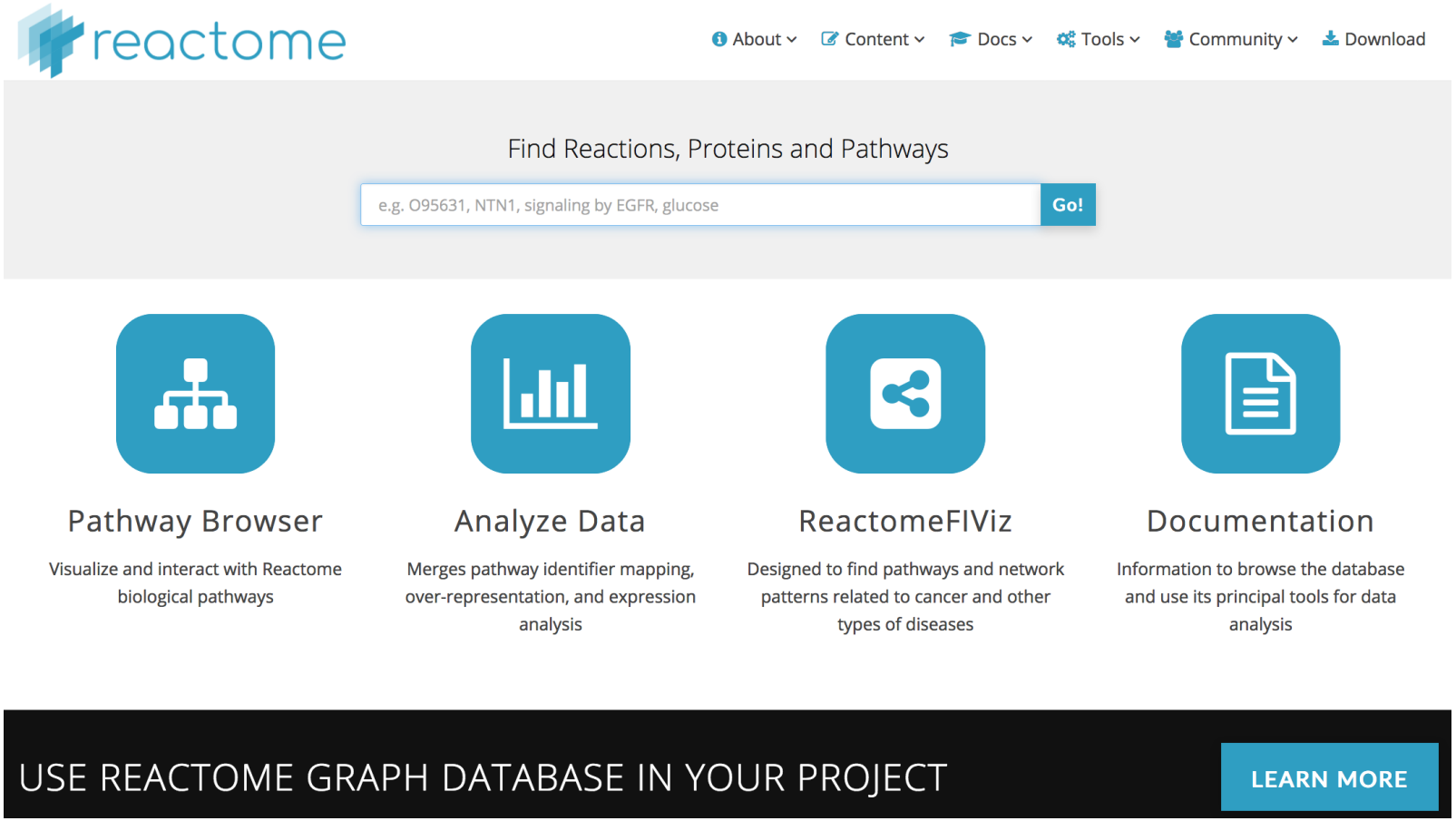
http://www.wikipathways.org

Wikipathways: apoptosis pathway



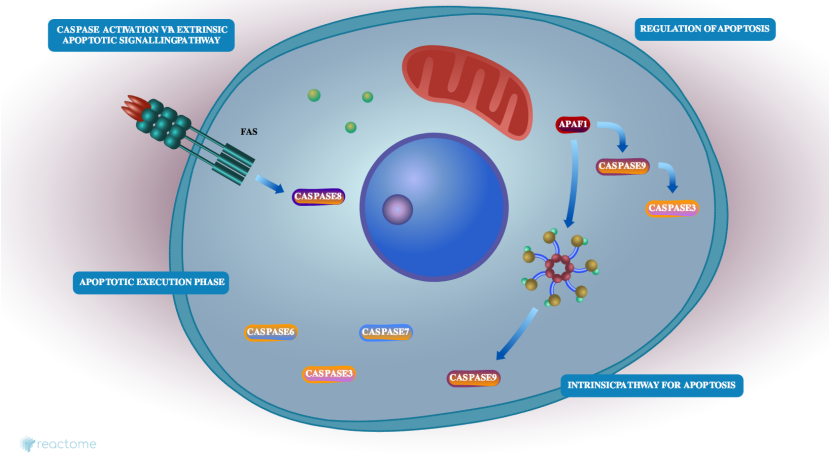
http://www.wikipathways.org/index.php/Pathway:WP254

User-curated database: Reactome



http://www.reactome.org

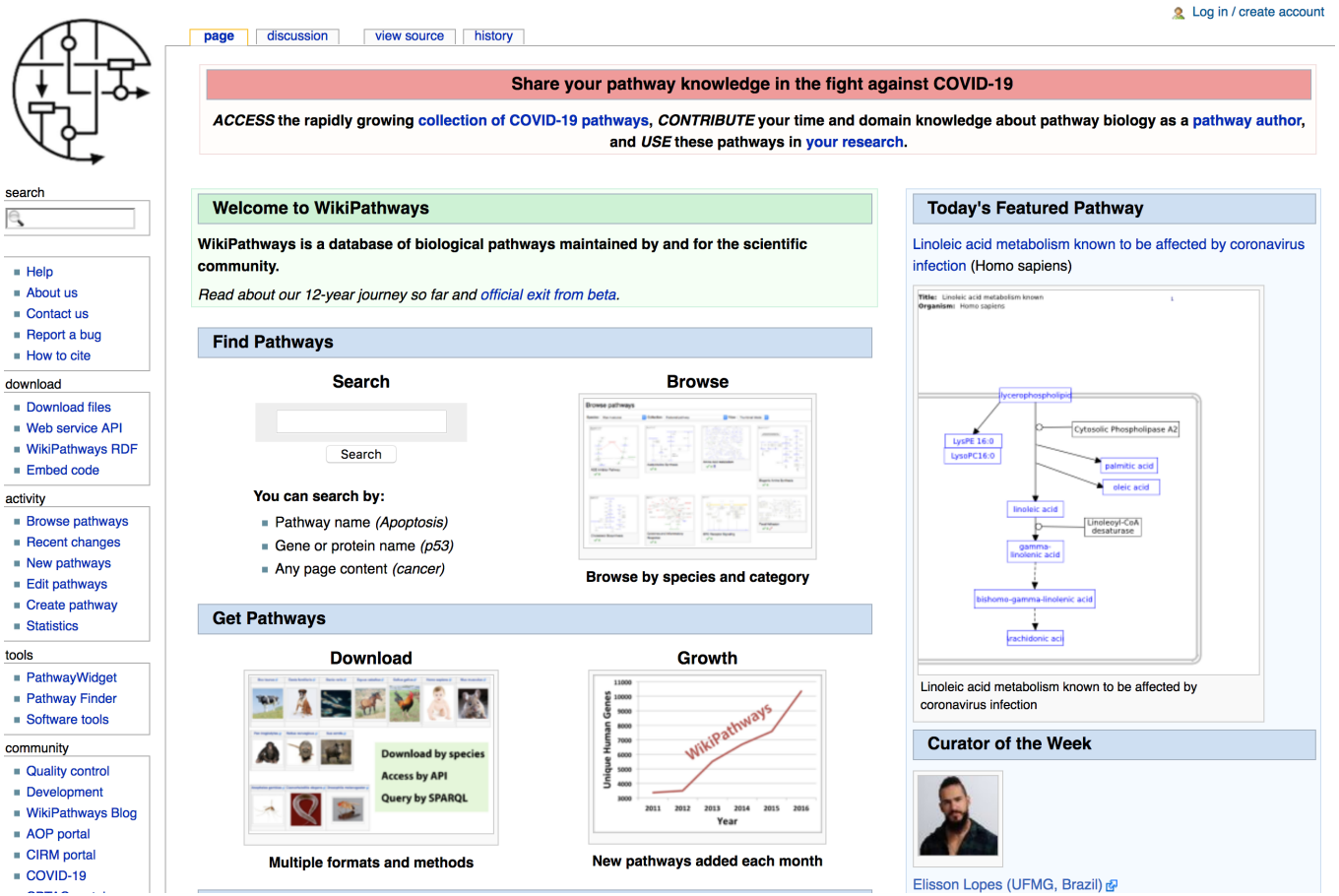
Reactome: apoptosis pathway



- The interactive pathway browser lets you explore the compone within each pathway.
- A Bioconductor package exists that lists the genes involved in ea Reactome pathway (`reactome.db`) - allows pathway information be incorporated into visualisation and analysis.

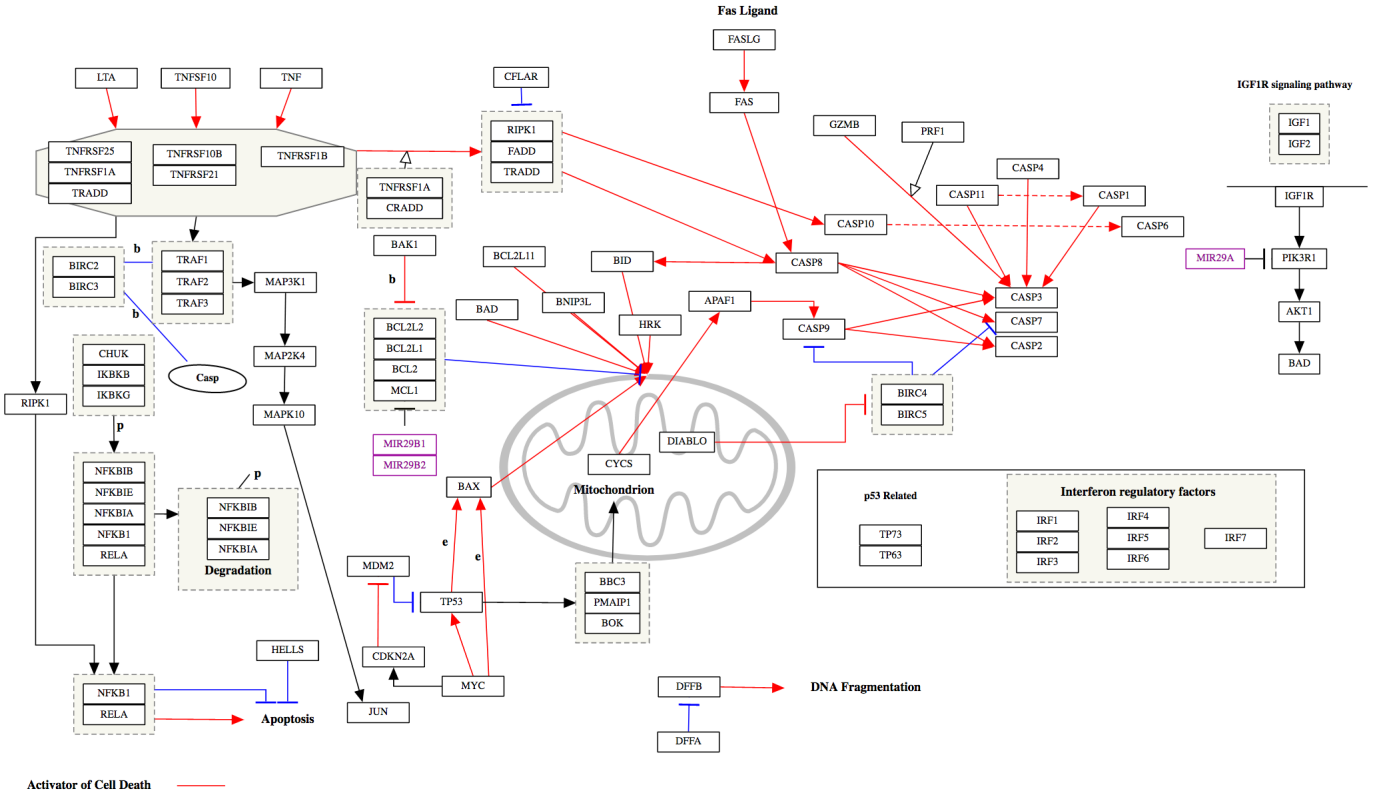
https://reactome.org/PathwayBrowser/#/R-HSA-109581

User-curated database: Wikipathways



http://www.wikipathways.org

Wikipathways: apoptosis pathway



https://www.wikipathways.org/index.php/Pathway:WP254

Gene Ontology

- Gene Ontology (GO) defines a collection of words (an ontolo which are used to classify the function of a gene.
- Three broad classifications:
  - Molecular function.
  - Biological process.
  - Cellular component.
- Each of these broad terms contains a hierarchy of categories, go from general to specific.
- Each category is indexed by an identifier.

Example of GO hierarchy (apoptosis)

```
*  all : all   ( 218850 )
        o GO:0008150 : biological_process ( 145098 )
              + GO:0009987 : cellular process ( 91236 )
                    # GO:0050875 : cellular physiological process (81383 )
                          * GO:0008219 : cell death ( 2714 )
                                o GO:0012501 : programmed cell death ( 2395 )
                                      + GO:0006915 : apoptosis ( 2061 )
              + GO:0007582 : physiological process ( 96419 )
                    # GO:0050875 : cellular physiological process ( 81383 )
                          * GO:0008219 : cell death ( 2714 )
                                o GO:0012501 : programmed cell death ( 2395 )
                                      + GO:0006915 : apoptosis ( 2061 )
                    # GO:0016265 : death ( 3054 )
                          * GO:0008219 : cell death ( 2714 )
                                o GO:0012501 : programmed cell death ( 2395 )
                                      + GO:0006915 : apoptosis ( 2061 )
```

Annotation for transcriptomics

- Linking information back to the transcript fragments.
- Types of information:
  - Sequence.
  - Gene.
  - Chromosome location.
  - Publications.
  - Function.
  - Other (e.g., transcription factors, orthologs, proteins).
- Amount of information available is organism-specific.

Common identifiers

- Microarray experiments often utilize the following identifiers:
  - Manufacturer's ID (e.g., Affymetrix probe ID).
  - Accession number (sequence identifier).
- RNA-seq and microarray experiments often include:
  - Entrez Gene ID or Ensembl ID (gene identifier).
  - KEGG ID (KEGG pathway membership).
  - GO term ID (functional information).
- This information can be used to enhance (or be incorporated intc statistical analysis.

Annotation in Bioconductor

- Bioconductor includes metadata packages which cont annotation information.
  - Array specific (e.g., Affymetrix HGU133A).
  - Organism specific (e.g., human, rat, mouse).
  - Database specific (e.g., GO, Reactome,)
- These packages provide linkage between the sequences used transcriptomic experiments, and the genes from which they a derived.
- GO and KEGG (and other) libraries are also available, with links Entrez Gene IDs.

Detecting pathway-level changes

- Transcriptomic experiments are able to measure changes in ge expression across treatment conditions.
- Can obtain information about gene sets (e.g., GO, KEGG, Reactom
- Allows transcriptomic data to be used to assess whether changes expression occur at the group level.
- Such changes often provide greater information than single ge changes.

Hypergeometric distribution

- Simple approach to investigating coordinated gene expressio involves hypergeometric distribution.
- Look for functional groupings within a set of significar differentially expressed genes:
  - e.g., what is the probability of getting 10 apoptosis genes in 100 differentially expressed genes?
- Similar to classic hypergeometric problem:
  - e.g., what is the probability of selecting k white balls ir sample of size n from a bag containing m white and $N -$ black balls?

Hypergeometric distribution

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, x = \max(0, n + M - N) \text{ to } x = \min(n, M)$$

- Here x is the number of genes from a particular pathway (of size which showed up in our list of n differentially expressed ger (then are N genes in total).
- To calculate a p-value for this "test" we need to sum up all of t probabilities from x (which we observed) up to $min(M, n)$.
- This is done for each gene set, and then the p-values are adjust to take multiple comparisons into account.

Fisher's Exact Test

- In practice we can use Fisher's Exact Test to determine whethe functional grouping is over-represented (or enriched) in our list differentially expressed genes.
  - This is a test for independence in a 2×2 table.
- Suppose that we observe 10 apoptosis genes in our 1 differentially expressed genes, and there are 10,000 genes on c array, of which 500 are apoptosis genes.
- Fisher's Exact Test uses the hypergeometric distribution to t whether being involved in apoptosis is independent of be significantly differentially expressed in our hypothetical experime

How would we do this in R?

```
## Create a matrix representing our data
x <- matrix(c(10,490,90,9410),2,2)
x
```

```
##      [,1] [,2]
## [1,]   10   90
## [2,]  490 9410
```

```
## Row and column sums
colSums(x)
```

```
## [1]  500 9500
```

```
rowSums(x)
```

Test for association

```
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.03328
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.9832142 4.1416491
## sample estimates:
## odds ratio
##   2.133664
```

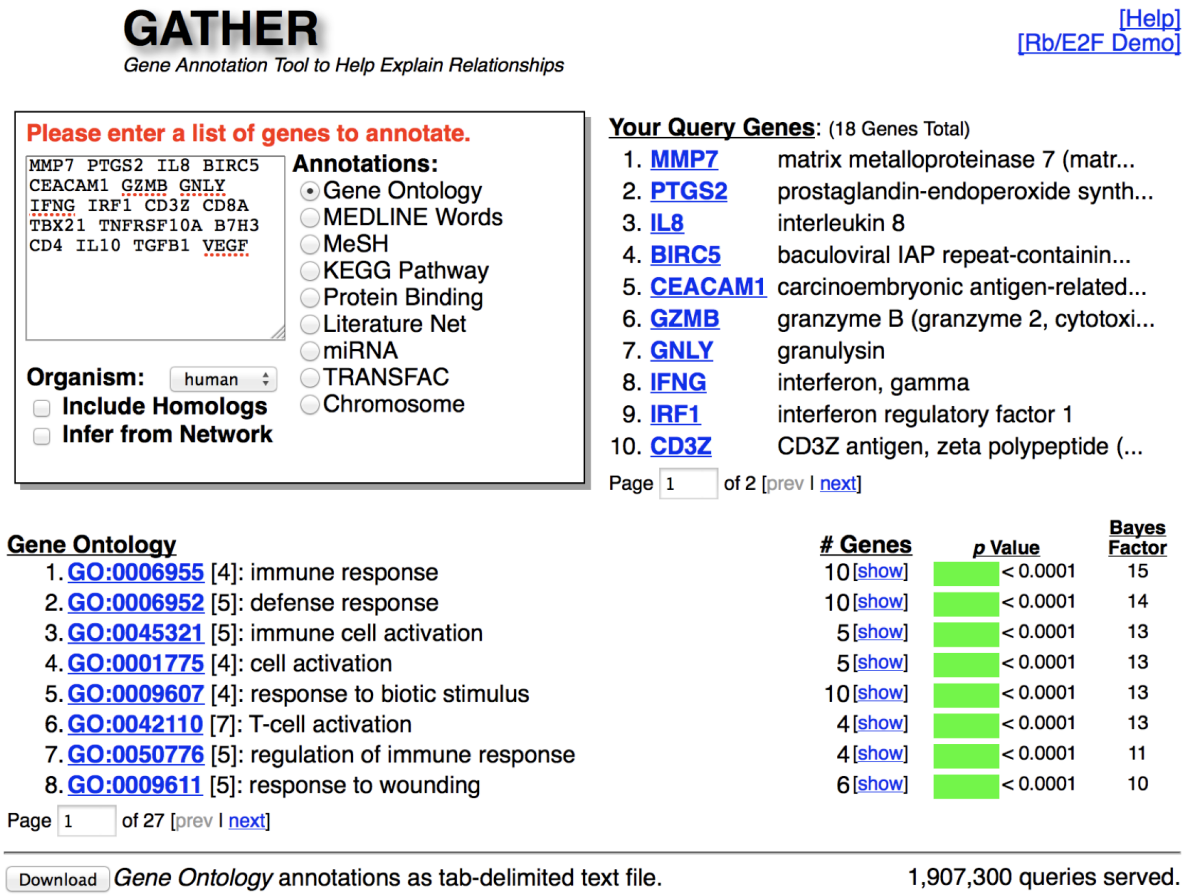
Tools for over-representation analysis

- There are MANY R-based and online tools for assessing functio enrichment of gene lists.
- We'll look at two (slightly old ones) here: GATHER and GeneSetDB
- Two newer online resources that are worth checking out are:
  - PANTHER: http://pantherdb.org/
  - Enrichr: http://amp.pharm.mssm.edu/Enrichr/
- PANTHER also has a Bioconductor annotation package availa (PANTHER.db):
  - https://bioconductor.org/packages/release/data/annotation/html/PANTHER.db.html

Start with a list of genes

MMP7	matrix metalloproteinase 7
PTGS2	prostaglandin-endoperoxide synthase 2
IL8	interleukin 8
BIRC5	baculoviral IAP repeat-containing 5
CEACAM1	carcinoembryonic antigen-related cell adhesion molecule 1
GZMB	granzyme B
GNLY	granulysin
IFNG	interferon, gamma
IRF1	interferon regulatory factor 1
CD3Z	CD3Z antigen, zeta polypeptide
CD8A	CD8 antigen, alpha polypeptide
TBX21	T-box 21
TNFRSF10A	tumor necrosis factor receptor superfamily, member 10a
B7H3	B7 homolog 3
CD4	CD4 antigen (p55)
IL10	interleukin 10
TGFB1	transforming growth factor, beta 1
VEGF	vascular endothelial growth factor

GATHER: Gene Ontology (GO)



http://gather.genome.duke.edu

GATHER: KEGG pathways

GATHER

Gene Annotation Tool to Help Explain Relationships

[Help]
[Rb/E2F Demo]

Please enter a list of genes to annotate.

```
MMP7 PTGS2 IL8 BIRC5
CEACAM1 GZMB GNLY
IFNG IRF1 CD3Z CD8A
TBX21 TNFRSF10A B7H3
CD4 IL10 TGFB1 VEGF
```

Annotations:

- ○ Gene Ontology
- ○ MEDLINE Words
- ○ MeSH
- ◉ KEGG Pathway
- ○ Protein Binding
- ○ Literature Net
- ○ miRNA
- ○ TRANSFAC
- ○ Chromosome

Organism: human

- ☐ **Include Homologs**
- ☐ **Infer from Network**

Your Query Genes: (18 Genes Total)

1. MMP7 matrix metalloproteinase 7 (matr...
2. PTGS2 prostaglandin-endoperoxide synth...
3. IL8 interleukin 8
4. BIRC5 baculoviral IAP repeat-containin...
5. CEACAM1 carcinoembryonic antigen-related...
6. GZMB granzyme B (granzyme 2, cytotoxi...
7. GNLY granulysin
8. IFNG interferon, gamma
9. IRF1 interferon regulatory factor 1
10. CD3Z CD3Z antigen, zeta polypeptide (...

Page 1 of 2 [prev | next]

KEGG Pathway	# Genes	*p* Value	Bayes Factor
1. path:hsa04060: Cytokine-cytokine receptor interaction IFNG IL10 IL8 TGFB1 TNFRSF10A VEGF	6 [hide]	< 0.0001	10
2. path:hsa04350: TGF-beta signaling pathway	2 [show]	0.002	2
3. path:hsa04630: Jak-STAT signaling pathway	2 [show]	0.007	1
4. path:hsa00590: Prostaglandin and leukotriene metabolism	1 [show]	0.01	0

http://gather.genome.duke.edu

GATHER: detailed output (GO results)

#	Annotation	Total Gen	Your Gene	Your Gene	Genome ('	Genome (	ln(Bayes f	neg ln(p v	FE: neg ln(	FE: neg ln(	Genes
1	GO:0006955	10	10	8	737	11524	14.91	11.17	17.81	12.13	B7H3 CD4 CD8A CEACAM1 GNLY IFNG IL10 IL8 IRF1 PTGS2
2	GO:0006952	10	10	8	828	11433	13.82	11.17	16.72	11.72	B7H3 CD4 CD8A CEACAM1 GNLY IFNG IL10 IL8 IRF1 PTGS2
3	GO:0045321	5	5	13	82	12179	12.95	11.17	15.88	11.55	B7H3 CD4 CD8A IL10 IL8
4	GO:0001775	5	5	13	83	12178	12.89	11.17	15.82	11.55	B7H3 CD4 CD8A IL10 IL8
5	GO:0009607	10	10	8	948	11313	12.56	11.17	15.45	11.55	B7H3 CD4 CD8A CEACAM1 GNLY IFNG IL10 IL8 IRF1 PTGS2
6	GO:0042110	4	4	14	32	12229	12.56	11.17	15.5	11.55	B7H3 CD4 CD8A IL10
7	GO:0050776	4	4	14	53	12208	10.68	11.17	13.62	9.88	B7H3 CD4 IL10 PTGS2
8	GO:0009611	6	6	12	280	11981	10.11	11.17	13.02	9.44	B7H3 CD4 GNLY IL10 IL8 PTGS2
9	GO:0042088	3	3	15	16	12245	9.95	11.17	12.89	9.44	B7H3 CD4 IL10
10	GO:0046649	4	4	14	67	12194	9.81	11.17	12.73	9.44	B7H3 CD4 CD8A IL10
11	GO:0042087	3	3	15	17	12244	9.79	11.17	12.73	9.44	B7H3 CD4 IL10
12	GO:0016066	3	3	15	20	12241	9.35	11.17	12.29	9.09	B7H3 CD4 IL10
13	GO:0006928	5	5	13	177	12084	9.3	11.17	12.21	9.09	IFNG IL10 IL8 PTGS2 VEGF
14	GO:0006968	4	4	14	83	12178	8.99	11.17	11.92	8.94	B7H3 CD4 GNLY IL10
15	GO:0050794	8	8	10	784	11477	8.99	11.17	11.86	8.94	B7H3 BIRC5 IFNG IL10 IL8 TGFB1 TNFRSF10A VEGF
16	GO:0042035	3	3	15	23	12238	8.97	11.17	11.91	8.94	B7H3 CD4 IL10
17	GO:0051244	7	7	11	560	11701	8.76	11.17	11.64	8.93	B7H3 BIRC5 IL10 IL8 TGFB1 TNFRSF10A VEGF
18	GO:0001817	3	3	15	25	12236	8.74	11.17	11.68	8.93	B7H3 CD4 IL10
19	GO:0051239	4	4	14	91	12170	8.64	11.17	11.57	8.93	B7H3 CD4 IL10 PTGS2
20	GO:0042089	3	3	15	26	12235	8.64	11.17	11.57	8.93	B7H3 CD4 IL10
21	GO:0042107	3	3	15	26	12235	8.64	11.17	11.57	8.93	B7H3 CD4 IL10
22	GO:0050918	2	2	16	2	12259	8.38	11.17	11.32	8.86	IL8 VEGF
23	GO:0050926	2	2	16	2	12259	8.38	11.17	11.32	8.86	IL8 VEGF
24	GO:0050927	2	2	16	2	12259	8.38	11.17	11.32	8.86	IL8 VEGF
25	GO:0050930	2	2	16	2	12259	8.38	11.17	11.32	8.86	IL8 VEGF
26	GO:0001816	3	3	15	29	12232	8.33	11.17	11.27	8.86	B7H3 CD4 IL10
27	GO:0043066	4	4	14	99	12162	8.32	11.17	11.25	8.86	BIRC5 IL10 TGFB1 VEGF
28	GO:0043069	4	4	14	100	12161	8.28	11.17	11.21	8.85	BIRC5 IL10 TGFB1 VEGF
29	GO:0050896	11	11	7	1933	10328	8.18	11.17	11.02	8.7	B7H3 CD4 CD8A CEACAM1 GNLY IFNG IL10 IL8 IRF1 PTGS2 VEGF
30	GO:0050920	2	2	16	3	12258	7.87	11.17	10.81	8.55	IL8 VEGF

http://gather.genome.duke.edu

GATHER: explanation of output

#	An index to number the annotations.
Annotation	The name of the annotation.
Total Genes With Ann	The number of genes from your list that have the annotation. If the *Include Homologs* inference is used, then this number will also include the homologous genes with the annotation from other organisms.
Your Genes (With Ann)	The number of genes from your list *with* the annotation.
Your Genes (No Ann)	The number of genes from your list *without* the annotation.
Genome (With Ann)	The number of genes in the genome (excluding those in your list) *with* the annotation.
Genome (No Ann)	The number of genes in the genome (excluding those in your list) *without* the annotation.
ln(Bayes factor)	The Bayes factor quantifying the amount of evidence supporting the hypothesis that the annotation is associated with your gene list. This is the same number that is shown on the website, but here, it is not rounded -- it contains more significant digits.
neg ln(p value)	The negative logarithm of the *p* value calculated from the Bayes factor (see Supplementary materials for the publication). The website shows the actual *p* values, but here, they are reported as logarithms for a more compact representation.
FE: neg ln(p value)	The negative logarithm of the *p* value calculated using a Fisher's exact test.
FE: neg ln(FDR)	The false discovery rate based on the Fisher's exact *p* value.
Genes	The symbols of the genes that have the annotation. If the *Include Homologs* inference is used, the homologous genes that have the annotation will also appear, but with a *:H* suffix. Similarly, if the *Infer from Network* inference is used, the genes that were included based on the network inference, and also have the annotation, will have a *:N* suffix.

http://gather.genome.duke.edu

Let's check (part of) the first row in R...

```
fisher.test(matrix(c(10,737,8,11524),2,2))
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  matrix(c(10, 737, 8, 11524), 2, 2)
## p-value = 1.833e-08
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   6.917398 57.039187
## sample estimates:
## odds ratio
##   19.53031
```

Let's check (part of) the first row in R...

```
-log(fisher.test(matrix(c(10,737,8,11524),2,2))$p.value)
```

```
## [1] 17.81479
```

GeneSetDB: A comprehensive meta-database, statistical and visualisation framework for gene set analysis

Open Access Article

Hiromitsu Araki[a], Christoph Knapp[b], Peter Tsai[b], Cristin Print[a, b]

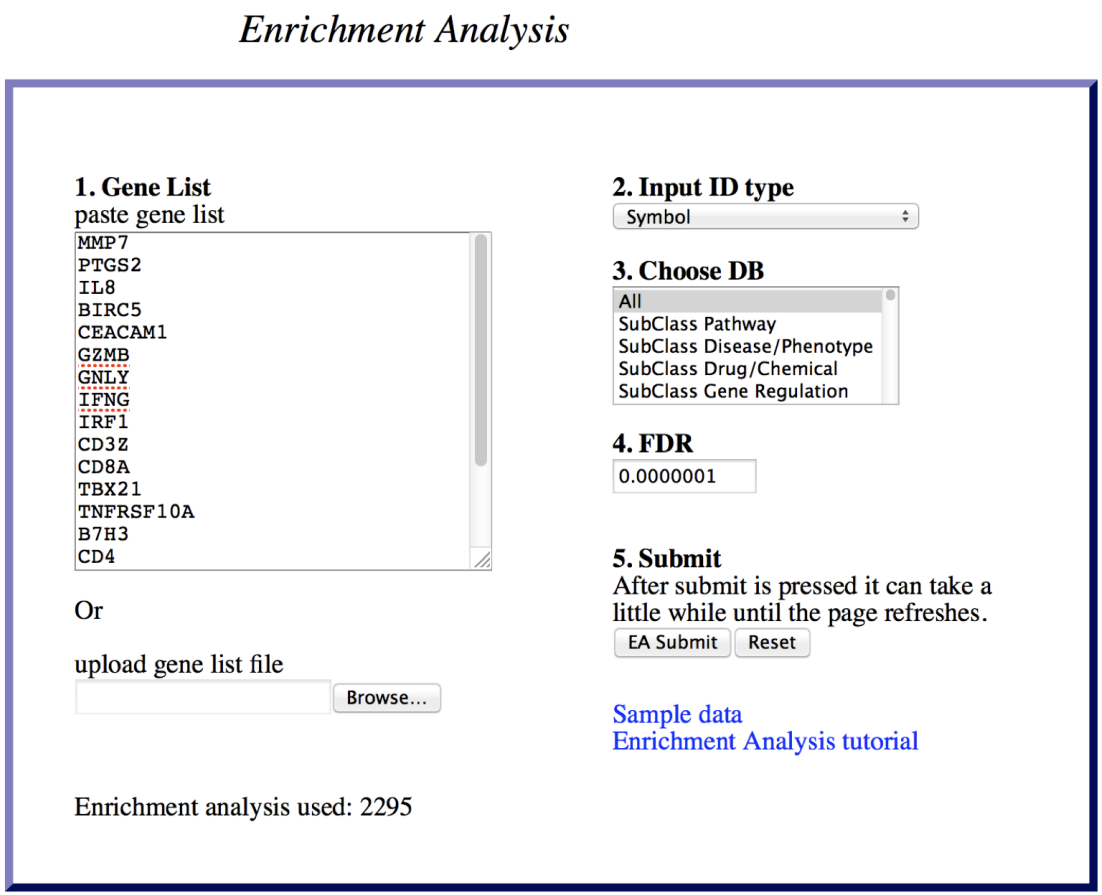
[a] Department of Molecular Medicine & Pathology, School of Medical Sciences, Faculty of Medical and Health Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand

[b] Bioinformatics Institute, The University of Auckland, Private Bag 92019, Auckland, New Zealand

http://dx.doi.org/10.1016/j.fob.2012.04.003, How to Cite or Link Using DOI

http://genesetdb.auckland.ac.nz

GeneSetDB: input



http://genesetdb.auckland.ac.nz

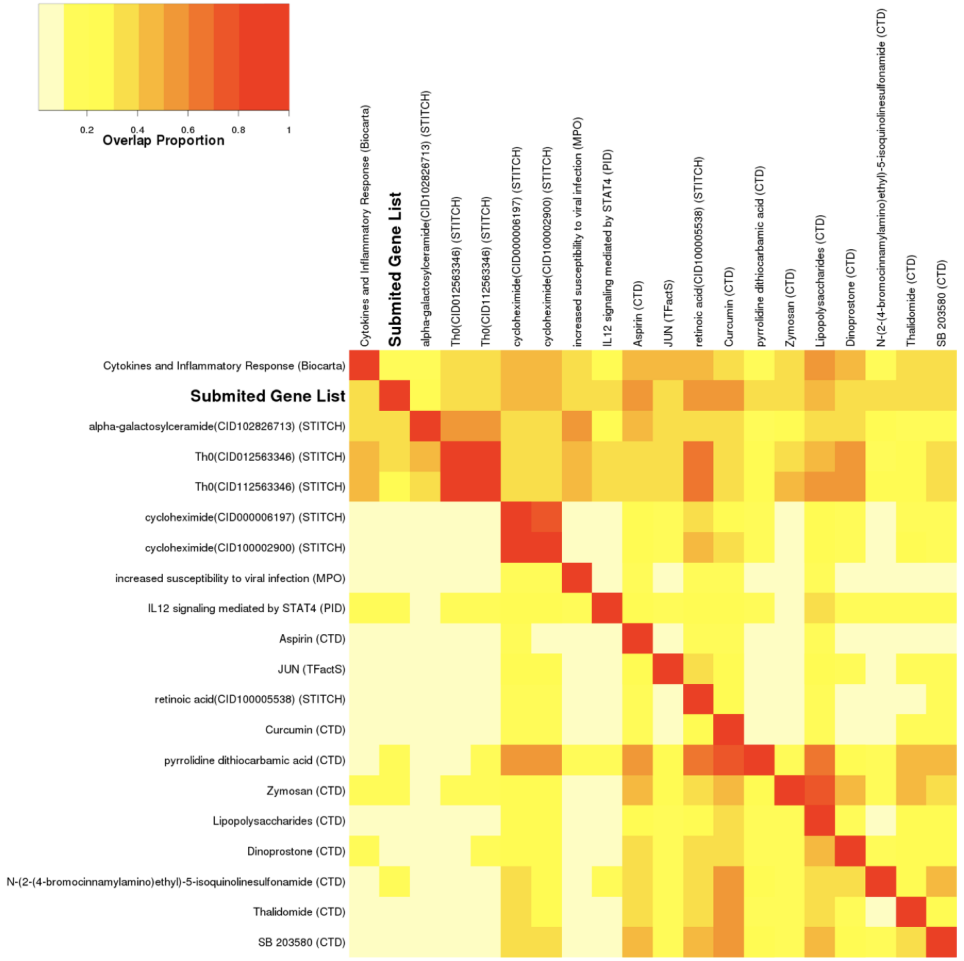
GeneSetDB: output

18 symbol input ids were converted into 15 unique gene ids.
19 entries for fdr cutoff 0.0000001 estimated.

Sub Class	Gene Set Name	Source DB	Gene #	Gene # with Anno	Gene # without Anno	p-value	FD
Drug/Chemical	Th0(CID012563346)	STITCH	16	5	11	6.9E-13	
Drug/Chemical	Th0(CID112563346)	STITCH	18	5	13	1.4E-12	
Pathway	Cytokines and Inflammatory Response	Biocarta	26	5	21	1.0E-11	
Drug/Chemical	Aspirin	CTD	461	9	452	1.4E-11	
Drug/Chemical	Dinoprostone	CTD	90	6	84	5.1E-11	
Drug/Chemical	cycloheximide(CID100002900)	STITCH	179	7	172	4.0E-11	
Pathway	IL12 signaling mediated by STAT4	PID	35	5	30	5.1E-11	
Disease/Phenotype	increased susceptibility to viral infection	MPO	93	6	87	6.3E-11	
Drug/Chemical	Zymosan	CTD	39	5	34	9.0E-11	
Drug/Chemical	Lipopolysaccharides	CTD	204	7	197	1.0E-10	
Drug/Chemical	N-(2-(4-bromocinnamylamino)ethyl)-5-isoquinolinesulfonamide	CTD	42	5	37	1.3E-10	
Drug/Chemical	SB 203580	CTD	108	6	102	1.6E-10	
Drug/Chemical	alpha-galactosylceramide(CID102826713)	STITCH	13	4	9	1.9E-10	
Drug/Chemical	pyrrolidine dithiocarbamic acid	CTD	46	5	41	2.1E-10	
Drug/Chemical	cycloheximide(CID000006197)	STITCH	233	7	226	2.6E-10	
Drug/Chemical	Thalidomide	CTD	125	6	119	3.8E-10	
Drug/Chemical	retinoic acid(CID100005538)	STITCH	421	8	413	3.5E-10	
GeneRegulation	JUN	TFactS	125	6	119	3.8E-10	
Drug/Chemical	Curcumin	CTD	438	8	430	4.8E-10	

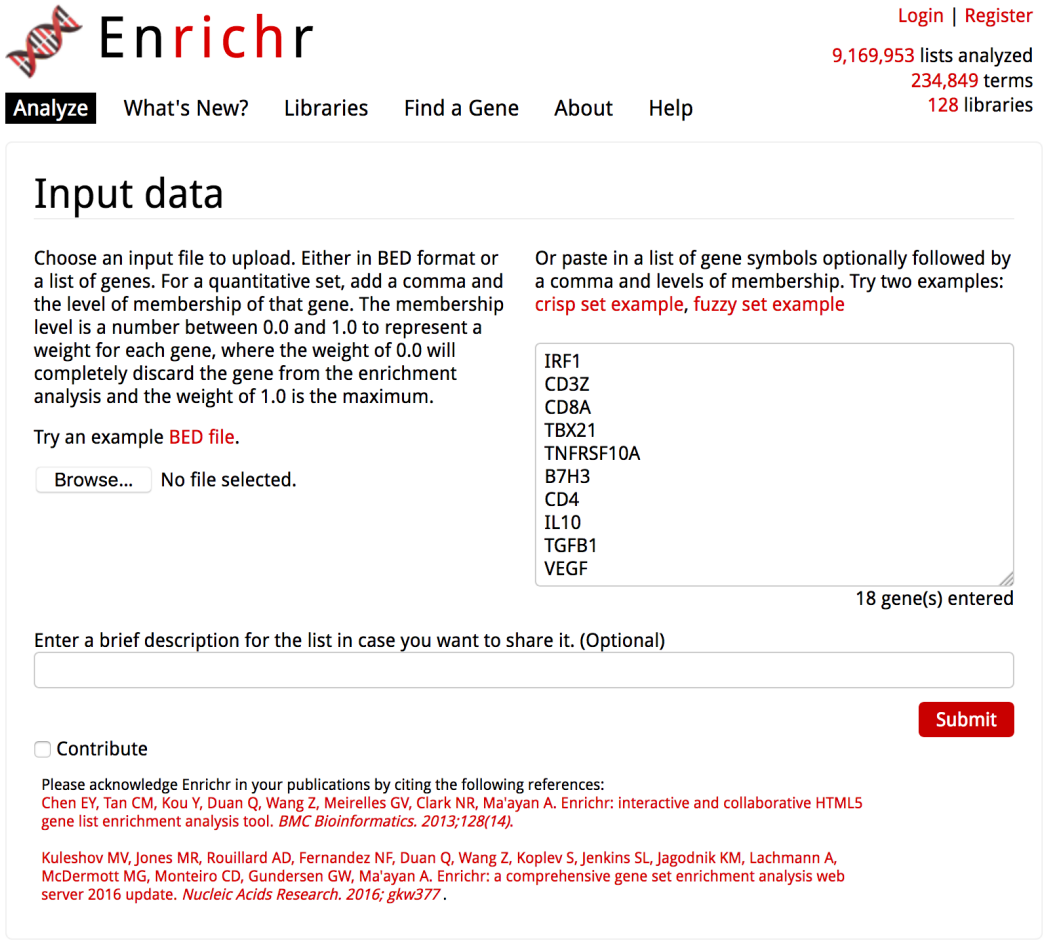
http://genesetdb.auckland.ac.nz

GeneSetDB: gene set overlap



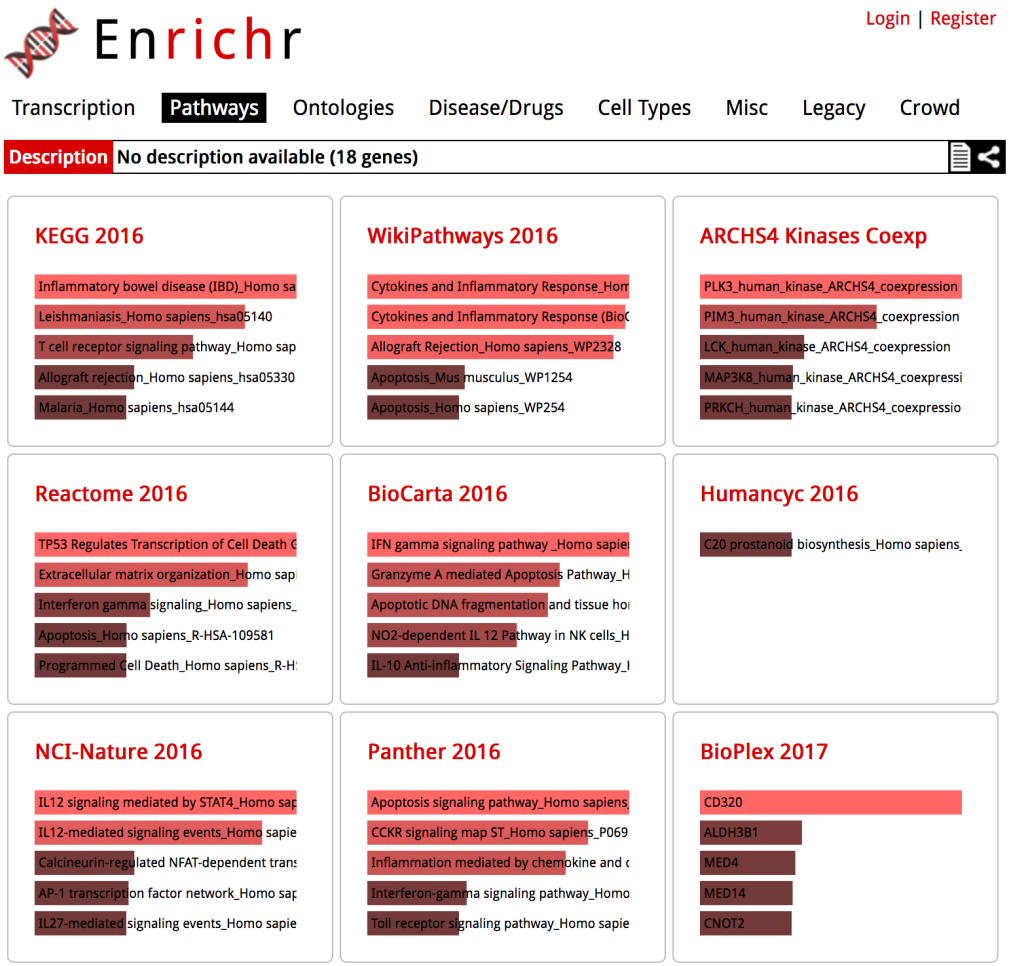
http://genesetdb.auckland.ac.nz

Enrichr



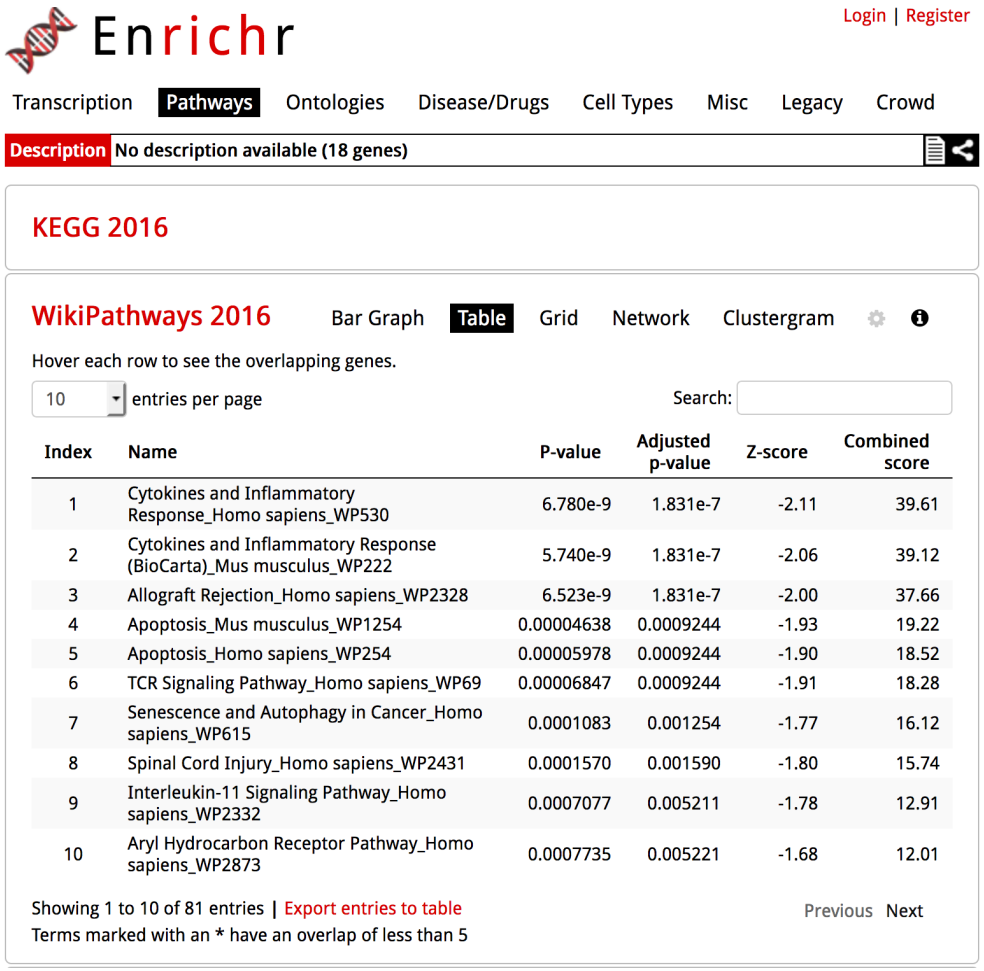
http://amp.pharm.mssm.edu/Enrichr/

Enrichr



http://amp.pharm.mssm.edu/Enrichr/

Enrichr



http://amp.pharm.mssm.edu/Enrichr/

Limitations of enrichment testing

- The hypergeometric-based enrichment tests only take the size gene sets into account.
- All genes for the same group that are not significant are treated t same.
  - What if they are "almost" significant?
  - We are now thinking about the ranks of the genes.
  - Can we incorporate this rank information into our calculatior
- Gene Set Enrichment Analysis (GSEA) provides a rank-bas assessment of enrichment, and doesn't require a list sigbnificantly differentially expressed genes.
- But that is a topic for another day...

Some caveats for RNA-seq data

- The gene-set analysis methods are applicable to transcriptor data from both microarrays and RNA-seq.
- One caveat, however, is that the results need to take gene leng into account.
  - RNA-seq tends to produce higher expression levels (i.e., grea counts) for longer genes: a longer transcript implies mo aligned fragments, and thus higher counts. This also giv these genes a great chance of being statistically differentia expressed.
  - Some gene sets (pathways, GO terms) tend to involve famil of long genes: if long genes have a great chance of be detected as differentially expressed, then gene sets consist of long genes will have a great chance of appeared to enriched in the analysis.