RNA-seq workshop

Annotation data and gene set analysis

Mik Black & Ngoni Faya
Genomics Aotearoa

Annotation

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

Entrez Gene

- Each putative gene in the genome is assigned an identifier in the "Entrez gene" database (and MANY other databases too).
- The gene identifier is also linked to a more descriptive gene name. This usually conveys some information about what that gene does (or at least what it was understood to be involved in at the time it was named).
- In transcriptomic experiments this means that we can find out the identity of genes that undergo differential expression.
- Depending on what is known about these genes, this information may provide important clues about the underlying biological process being studied.
- Although a gene name is often somewhat informative, vast amounts of information about that gene may reside in journal publications and internet databases - how do we get this information?

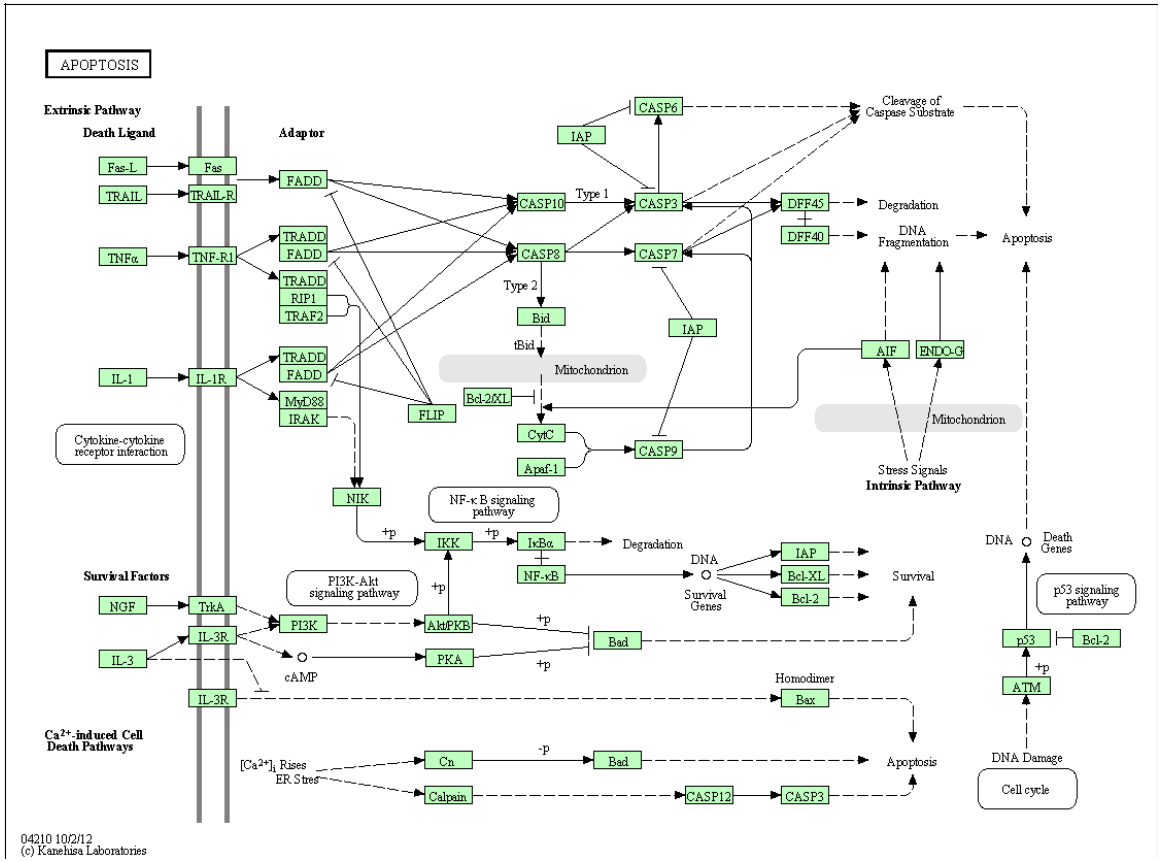
Biological pathways

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

Biological pathways: KEGG

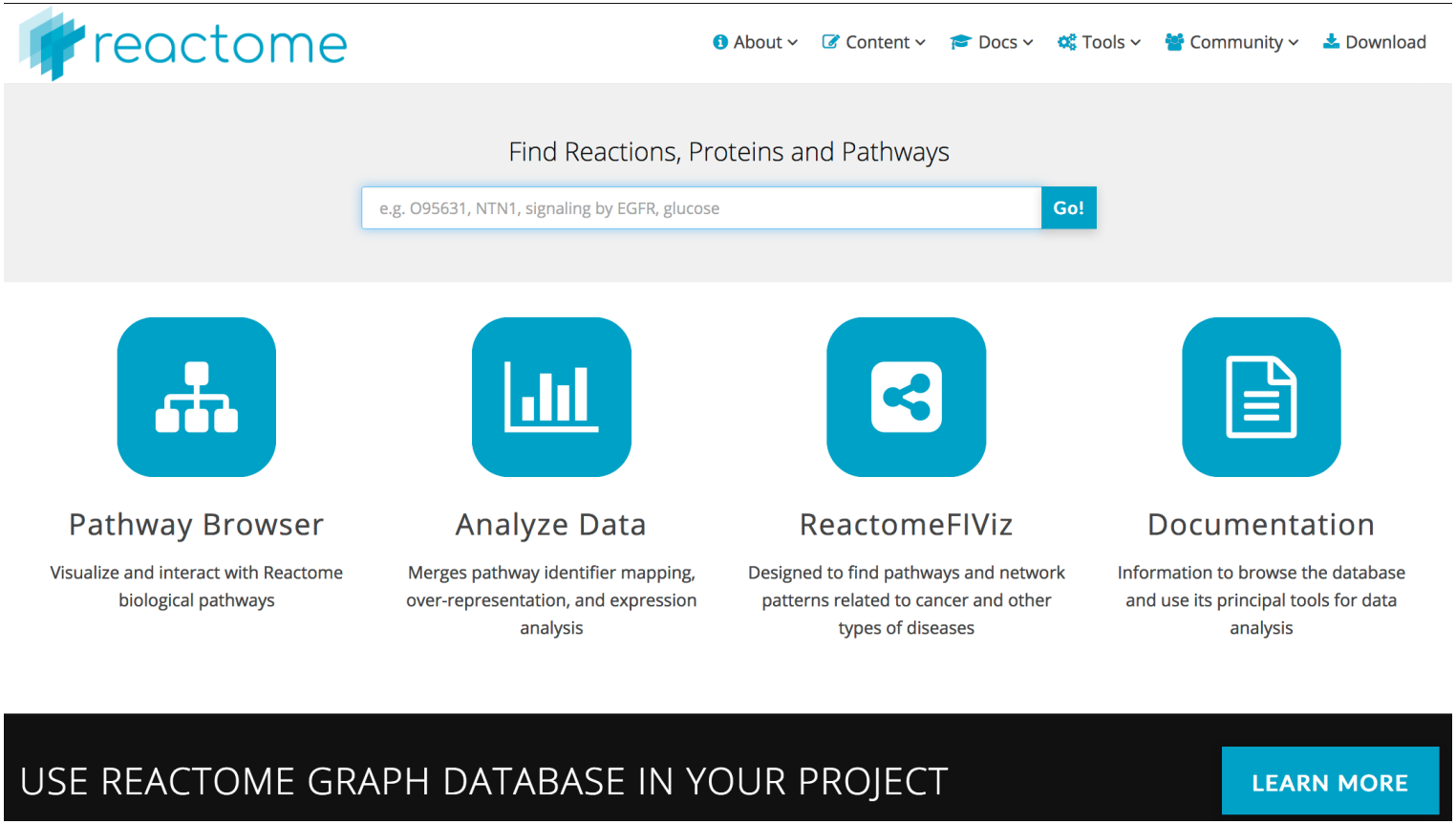
- Kyoto Encyclopedia of Gene and Genomes: http://www.genome.jp/kegg/kegg4.html
- Provides nice (user-created) pathway diagrams (although this is an old database, so the style of the diagrams looks a bit dated).
- XML output includes information about genes involved in pathways, and inter-gene (and gene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

KEGG pathway diagram (apoptosis)



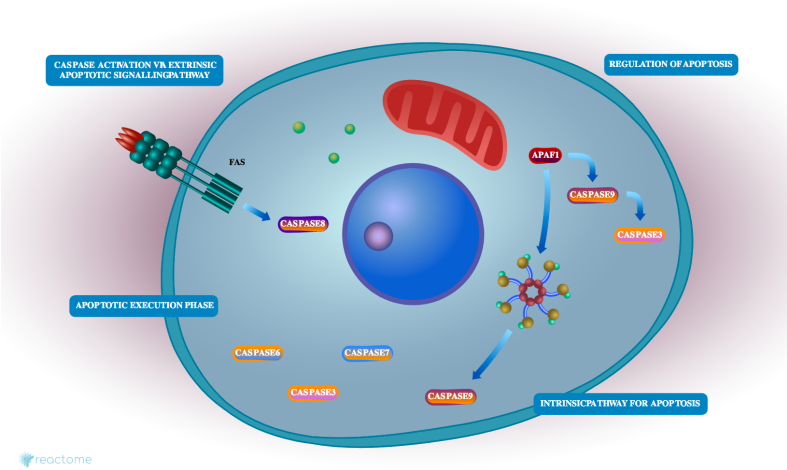
http://www.genome.jp/kegg-bin/show_pathway?org_name=hsa&mapno=04210&mapscale=&show_description=show

User-curated database: Reactome



http://www.reactome.org

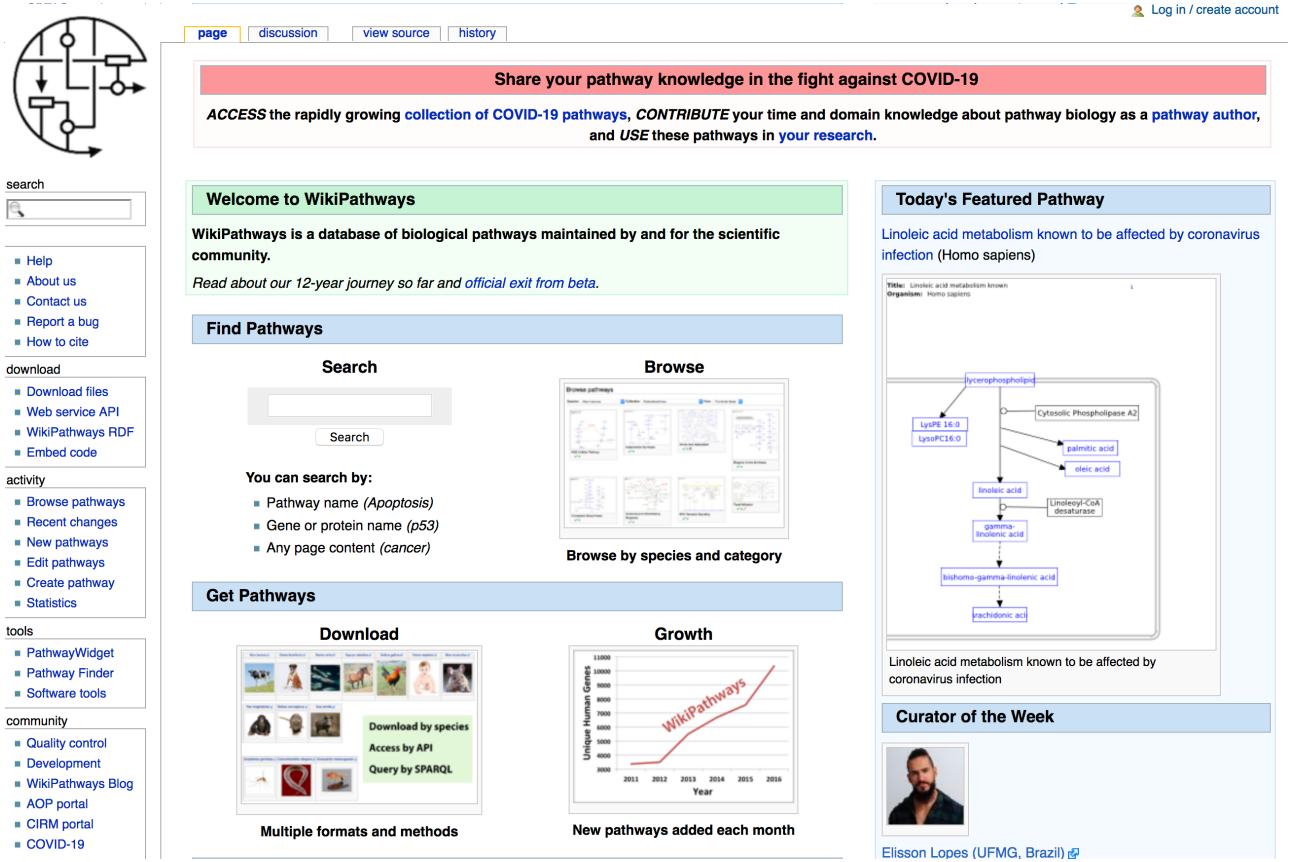
Reactome: apoptosis pathway



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

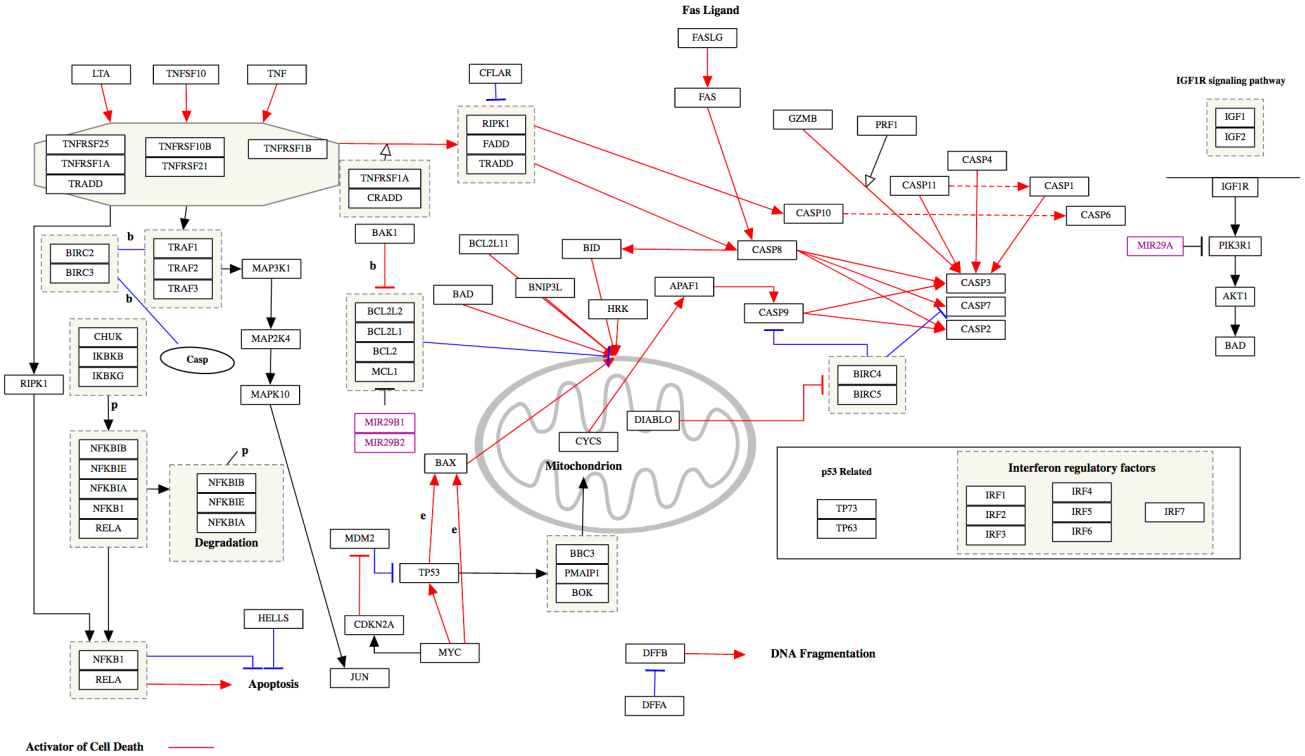
https://reactome.org/PathwayBrowser/#/R-HSA-109581

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 ontology) 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, going 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.

Annotation in Bioconductor

- Bioconductor includes metadata packages which contain annotation information.
  - Microarray 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 in transcriptomic experiments, and the genes from which they are derived.
- GO and KEGG (and other) libraries are also available, with links to Entrez Gene IDs.

Detecting pathway-level changes

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

Hypergeometric distribution

- Simple approach to investigating coordinated gene expression - involves hypergeometric distribution.
- Look for functional groupings within a set of significantly differentially expressed genes:
  - e.g., what is the probability of getting 10 apoptosis genes in my 100 differentially expressed genes?
- Similar to classic hypergeometric problem:
  - e.g., what is the probability of selecting k white balls in a sample of size n from a bag containing m white and $N - m$ 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 M) which showed up in our list of n differentially expressed genes (then are N genes in total).
- To calculate a p-value for this "test" we need to sum up all of the probabilities from x (which we observed) up to $min(M, n)$.
- This is done for each gene set, and then the p-values are adjusted to take multiple comparisons into account.

Fisher's Exact Test

- In practice we can use Fisher's Exact Test to determine whether a functional grouping is over-represented (or enriched) in our list of differentially expressed genes.
  - This is a test for independence in a 2×2 table.
- Suppose that we observe 10 apoptosis genes in our 100 differentially expressed genes, and there are 10,000 genes on our array, of which 500 are apoptosis genes.
- Fisher's Exact Test uses the hypergeometric distribution to test whether being involved in apoptosis is independent of being significantly differentially expressed in our hypothetical experiment.

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)
```

```
## [1]  100 9900
```

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
```

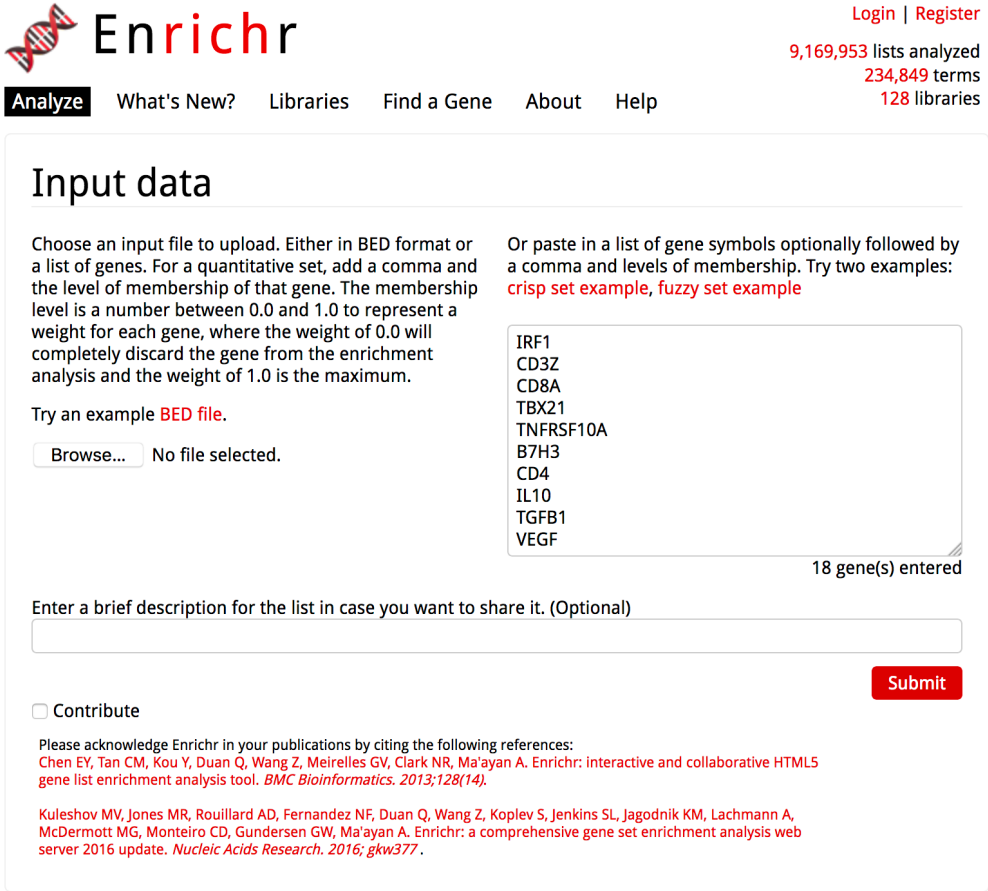
Tools for over-representation analysis

- There are MANY R-based and online tools for assessing functional 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 available (PANTHER.db):
  - https://bioconductor.org/packages/release/data/annotation/html/PANTHER.db.html

Start with a list of genes

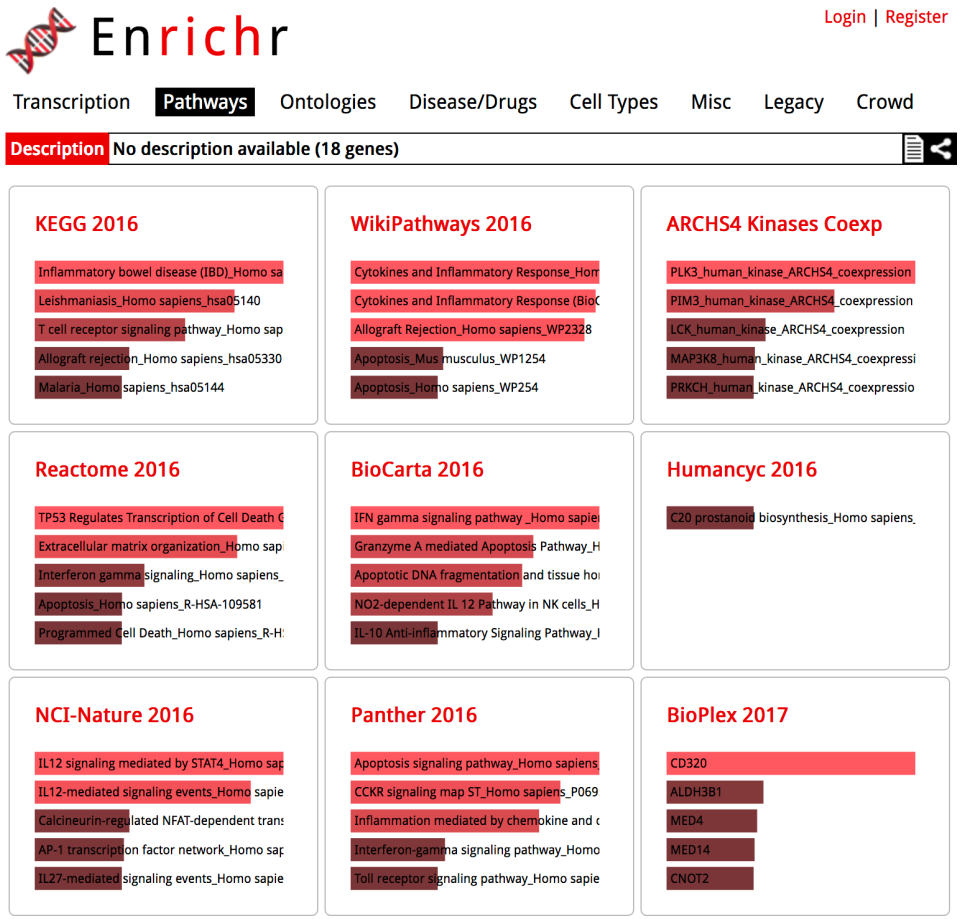
MMP7	matrix metalloproteinase 7
PTGS2	prostaglandin-endoperoxide synthase 2
IL8	interleukin 8
BIRC5	baculoviral IAP repeat-containing 5
CEACAM1	carcinoembryonic antigen-related cell adhesion molecule 1
GZMB	granzyme B
GNLY	granulysin
IFNG	interferon, gamma
IRF1	interferon regulatory factor 1
CD3Z	CD3Z 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

Enrichr



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

Enrichr



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

Enrichr

Enrichr

Login | Register

Transcription **Pathways** Ontologies Disease/Drugs Cell Types Misc Legacy Crowd

Description No description available (18 genes)

KEGG 2016

WikiPathways 2016

Bar Graph **Table** Grid Network Clustergram

Hover each row to see the overlapping genes.

10 entries per page

Search:

Index	Name	P-value	Adjusted p-value	Z-score	Combined score
1	Cytokines and Inflammatory Response_Homo sapiens_WP530	6.780e-9	1.831e-7	-2.11	39.61
2	Cytokines and Inflammatory Response (BioCarta)_Mus musculus_WP222	5.740e-9	1.831e-7	-2.06	39.12
3	Allograft Rejection_Homo sapiens_WP2328	6.523e-9	1.831e-7	-2.00	37.66
4	Apoptosis_Mus musculus_WP1254	0.00004638	0.0009244	-1.93	19.22
5	Apoptosis_Homo sapiens_WP254	0.00005978	0.0009244	-1.90	18.52
6	TCR Signaling Pathway_Homo sapiens_WP69	0.00006847	0.0009244	-1.91	18.28
7	Senescence and Autophagy in Cancer_Homo sapiens_WP615	0.0001083	0.001254	-1.77	16.12
8	Spinal Cord Injury_Homo sapiens_WP2431	0.0001570	0.001590	-1.80	15.74
9	Interleukin-11 Signaling Pathway_Homo sapiens_WP2332	0.0007077	0.005211	-1.78	12.91
10	Aryl Hydrocarbon Receptor Pathway_Homo sapiens_WP2873	0.0007735	0.005221	-1.68	12.01

Showing 1 to 10 of 81 entries | Export entries to table

Previous Next

Terms marked with an * have an overlap of less than 5

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

Limitations of enrichment testing

- The hypergeometric-based enrichment tests only take the size of gene sets into account.
- All genes for the same group that are not significant are treated the 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 calculations?
- Gene Set Enrichment Analysis (GSEA) provides a rank-based assessment of enrichment, and doesn't require a list of 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 transcriptomic data from both microarrays and RNA-seq.
- One caveat, however, is that the results need to take gene length into account.
  - RNA-seq tends to produce higher expression levels (i.e., greater counts) for longer genes: a longer transcript implies more aligned fragments, and thus higher counts. This also gives these genes a great chance of being statistically differentially expressed.
  - Some gene sets (pathways, GO terms) tend to involve families of long genes: if long genes have a great chance of being detected as differentially expressed, then gene sets consisting of long genes will have a great chance of appeared to be enriched in the analysis.