

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 related to sequences, genes, protein, function...

Gene annotation: the good old days...

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

Entrez Gene

- With the sequencing of the human genome, individual sequencer fragments could be mapped to their position in the genome, and annotated (in conjunction with other fragments) as "genes".
- Each putative gene in the genome was also assigned an identifier in the "Entrez gene" database (also at NCBI).
- The accession numbers of the constituent fragments were then associated with this identifier (and vice versa).
- 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).

Transcriptomics (microarrays and RNA-seq)

- Each spot on a microarray (or transcript fragment in an RNA-seq experiment) is associated with a gene (or some sort of meaningful DNA sequence), and can thus be linked to an accession number and/or Entrez ID.
- Through these numbers we can find out about the gene that the sequence is associated with (if known), where in the genome it is located, and (maybe) what it does.
- 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.

Gene function

- For biologists, it's not necessarily interesting to find out that a gene is significantly differentially expressed if no other information is known about that gene.
- One (very good) reason for this is that in transcriptomics experiments there are often a lot of false positives, so biologists 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 genes REALLY independent? How about the residuals - normally distributed?

Gene function

- If enough is known about a differentially expressed gene for it "make sense" or be "interesting" in the context of the experiment then biologists tend to get a bit more excited.
- Although a gene name is often somewhat informative, very large amounts of information about that gene may reside in journal publications and internet databases - how do we get this 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 back to the appropriate identifiers.
- Based on this the NCBI search engine hosts a local copy of the NCBI databases) can be used to retrieve information about differentially expressed genes.

Problem - too much information

- For situations where large numbers of genes are differentially 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 expressed genes which share a common function?

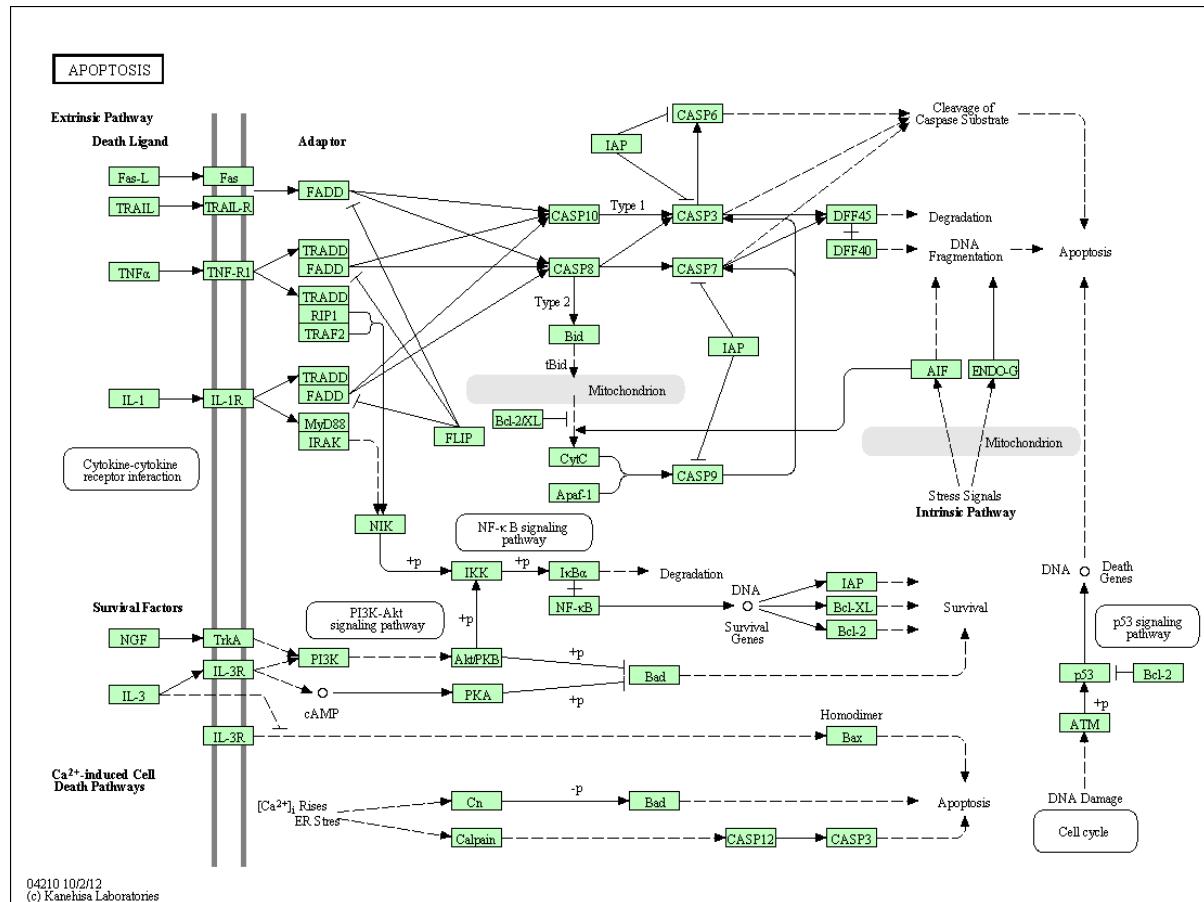
Biological pathways

- In reality genes are members of pathways, which perform many 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 is 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 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)



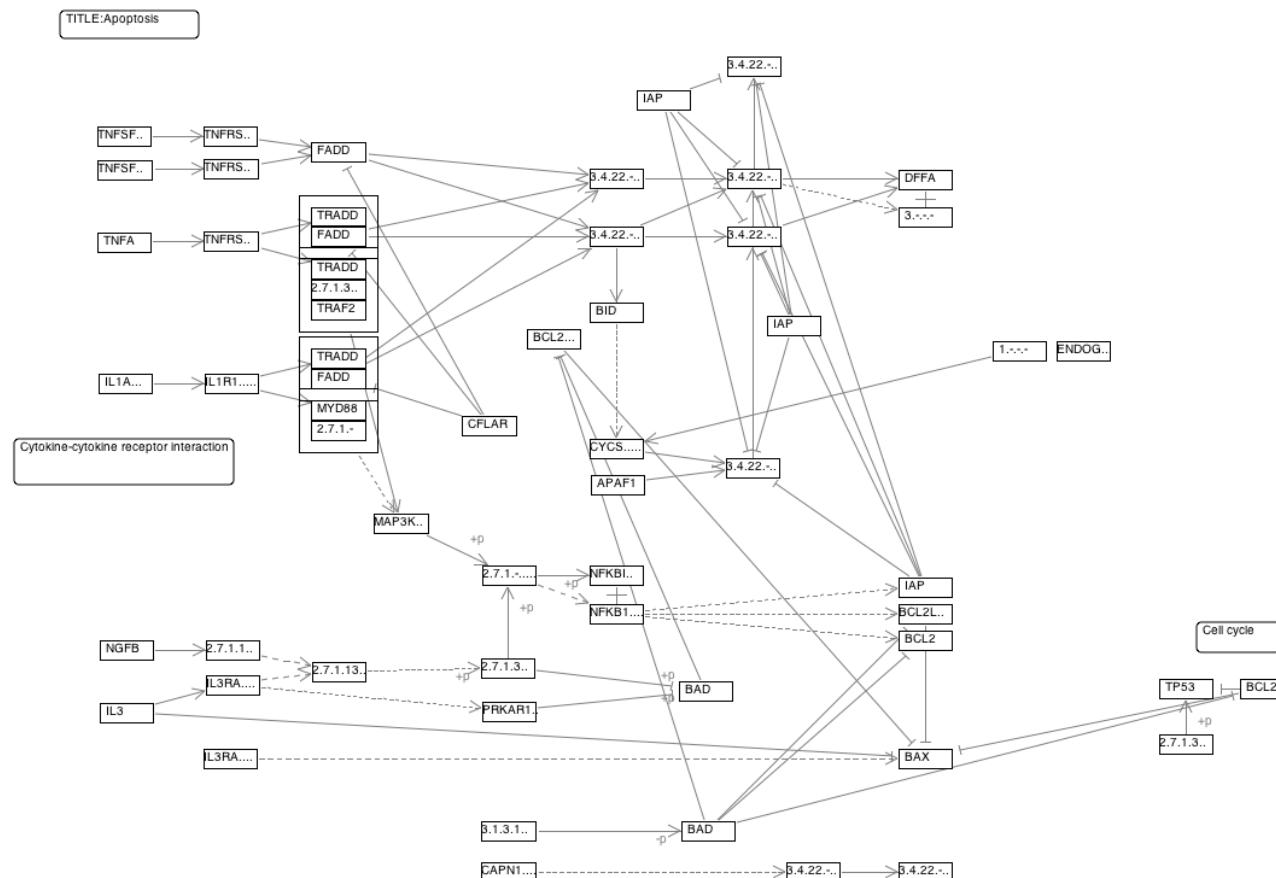
http://www.genome.jp/kegg-bin/show_pathway?org_name=hsa&mapno=04210&mapscale=&

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, 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

Share your pathway knowledge in the fight against COVID-19

ACCESS the rapidly growing collection of COVID-19 pathways, CONTRIBUTE your time and domain knowledge about pathway biology as a pathway author, and USE these pathways in your research.

Welcome to WikiPathways

WikiPathways is a database of biological pathways maintained by and for the scientific community.

Read about our 12-year journey so far and official exit from beta.

Find Pathways

Search

You can search by:

- Pathway name (*Apoptosis*)
- Gene or protein name (*p53*)
- Any page content (*cancer*)

Browse

Browse pathways

Browse by species and category

Get Pathways

Download

Multiple formats and methods

Growth

New pathways added each month

Today's Featured Pathway

Linoleic acid metabolism known to be affected by coronavirus infection (*Homo sapiens*)

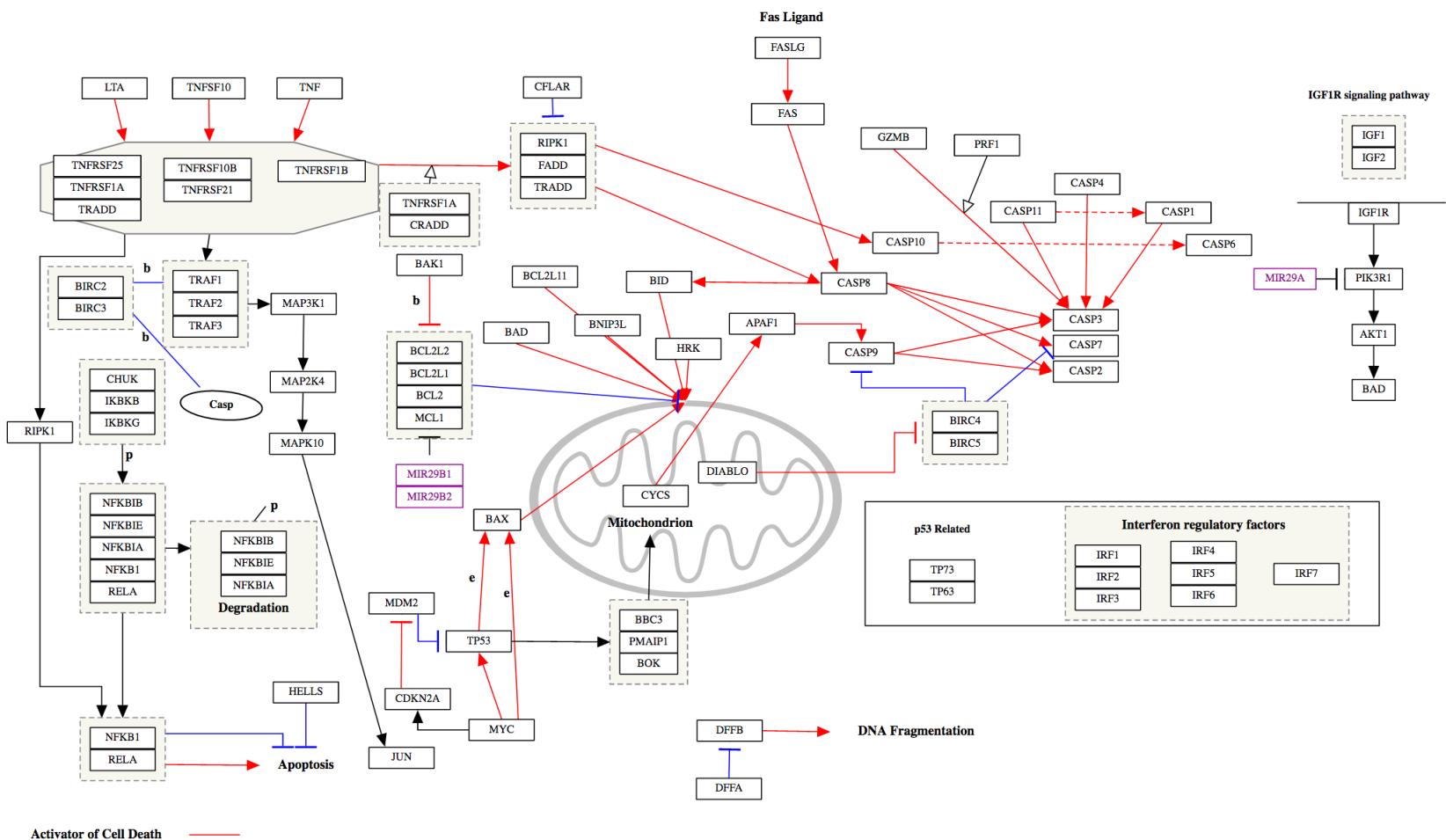
Title: Linoleic acid metabolism known
Organism: *Homo sapiens*

Curator of the Week

Elisson Lopes (UFMG, Brazil)

<http://www.wikipathways.org>

WikiPathways: apoptosis pathway



User-curated database: Reactome

The screenshot shows the Reactome website homepage. At the top is a navigation bar with the Reactome logo, followed by links for About, Content, Docs, Tools, Community, and Download. Below the navigation is a search bar with the placeholder "Find Reactions, Proteins and Pathways" and a "Go!" button. The main content area features four large blue icons with white symbols: a tree-like structure for the Pathway Browser, a bar chart for Analyze Data, a network graph for ReactomeFLViz, and a document for Documentation. Each icon has a descriptive title and a brief description below it. A black banner at the bottom encourages users to "USE REACTOME GRAPH DATABASE IN YOUR PROJECT" with a "LEARN MORE" button.

reactome

About Content Docs Tools Community Download

Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose Go!

Pathway Browser
Visualize and interact with Reactome biological pathways

Analyze Data
Merges pathway identifier mapping, over-representation, and expression analysis

ReactomeFLViz
Designed to find pathways and network patterns related to cancer and other types of diseases

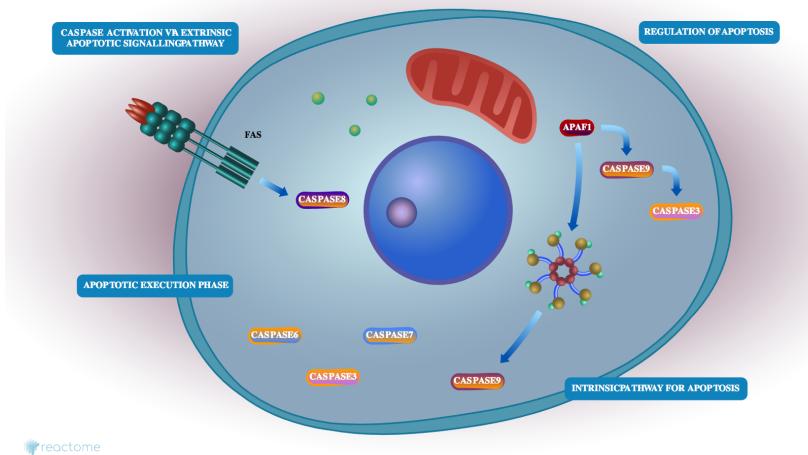
Documentation
Information to browse the database and use its principal tools for data analysis

USE REACTOME GRAPH DATABASE IN YOUR PROJECT

LEARN MORE

<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

Share your pathway knowledge in the fight against COVID-19

ACCESS the rapidly growing collection of COVID-19 pathways, CONTRIBUTE your time and domain knowledge about pathway biology as a pathway author, and USE these pathways in your research.

Welcome to WikiPathways

WikiPathways is a database of biological pathways maintained by and for the scientific community.

Read about our 12-year journey so far and official exit from beta.

Find Pathways

Search

You can search by:

- Pathway name (*Apoptosis*)
- Gene or protein name (*p53*)
- Any page content (*cancer*)

Browse

Browse pathways

Browse by species and category

Get Pathways

Download

Multiple formats and methods

Growth

New pathways added each month

Today's Featured Pathway

Linoleic acid metabolism known to be affected by coronavirus infection (*Homo sapiens*)

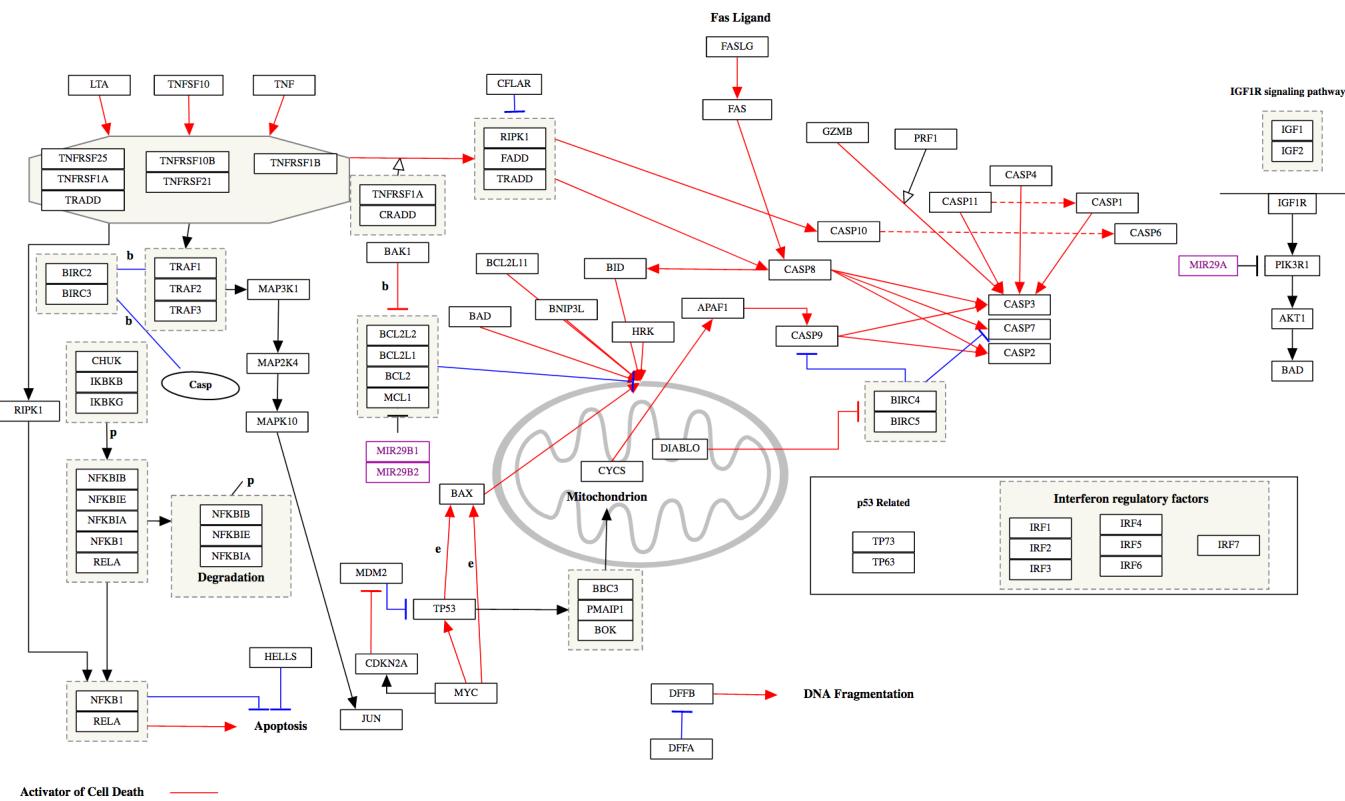
Title: Linoleic acid metabolism known
Organism: *Homo sapiens*

Curator of the Week

Elisson Lopes (UFMG, Brazil)

<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.

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 into) statistical analysis.

Annotation in Bioconductor

- Bioconductor includes metadata packages which contain 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 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 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 there 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 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 the 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)
```

```
...  
...
```

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 function 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

GATHER: Gene Ontology (GO)

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

Page of 2 [[prev](#) | [next](#)]

Gene Ontology

1. [GO:0006955](#) [4]: immune response
2. [GO:0006952](#) [5]: defense response
3. [GO:0045321](#) [5]: immune cell activation
4. [GO:0001775](#) [4]: cell activation
5. [GO:0009607](#) [4]: response to biotic stimulus
6. [GO:0042110](#) [7]: T-cell activation
7. [GO:0050776](#) [5]: regulation of immune response
8. [GO:0009611](#) [5]: response to wounding

Page of 27 [[prev](#) | [next](#)]

[Download](#) Gene Ontology annotations as tab-delimited text file.

1,907,300 queries served.

# Genes	p Value	Bayes Factor
10	< 0.0001	15
10	< 0.0001	14
5	< 0.0001	13
5	< 0.0001	13
10	< 0.0001	13
4	< 0.0001	13
4	< 0.0001	11
6	< 0.0001	10

<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

Organism: human
 Include Homologs
 Infer from Network

- Annotations:
- Gene Ontology
 - MEDLINE Words
 - MeSH
 - KEGG Pathway
 - Protein Binding
 - Literature Net
 - miRNA
 - TRANSFAC
 - Chromosome

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

1. [path:hsa04060](#): Cytokine-cytokine receptor interaction
IFNG IL10 IL8 TGFB1 TNFRSF10A VEGF
2. [path:hsa04350](#): TGF-beta signaling pathway
3. [path:hsa04630](#): Jak-STAT signaling pathway
4. [path:hsa00590](#): Prostaglandin and leukotriene metabolism

# Genes	p Value	Bayes Factor
6 [hide]	 < 0.0001	10
2 [show]	 0.002	2
2 [show]	 0.007	1
1 [show]	 0.01	0

<http://gather.genome.duke.edu>

GATHER: detailed output (GO results)

#	Annotation	Total Genes	Your Gene	Your Gene	Genome (-)	Genome (+)	In(Bayes f)	neg	In(p v)	FE: neg	In(FE: neg)	In(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 <i>Include Homologs</i> 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 <i>with</i> the annotation.
Your Genes (No Ann)	The number of genes from your list <i>without</i> the annotation.
Genome (With Ann)	The number of genes in the genome (excluding those in your list) <i>with</i> the annotation.
Genome (No Ann)	The number of genes in the genome (excluding those in your list) <i>without</i> 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 <i>Include Homologs</i> inference is used, the homologous genes that have the annotation will also appear, but with a :H suffix. Similarly, if the <i>Infer from Network</i> 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

 [Permissions & Reprints](#)

<http://genesetdb.auckland.ac.nz>

GeneSetDB: input

Enrichment Analysis

1. Gene List
paste gene list

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

Or

upload gene list file

2. Input ID type

3. Choose DB

- All
- SubClass Pathway
- SubClass Disease/Phenotype
- SubClass Drug/Chemical
- SubClass Gene Regulation

4. FDR

5. Submit
After submit is pressed it can take a little while until the page refreshes.

[Sample data](#)
[Enrichment Analysis tutorial](#)

Enrichment analysis used: 2295

<http://genesetdb.auckland.ac.nz>

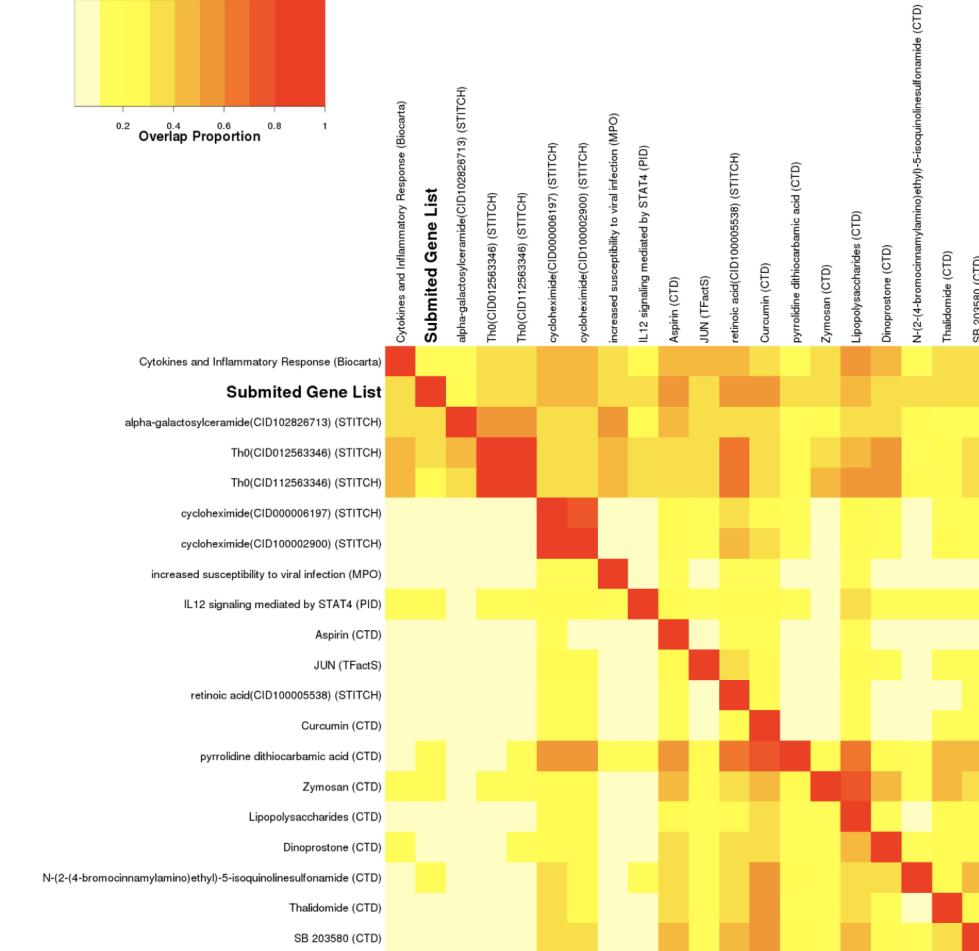
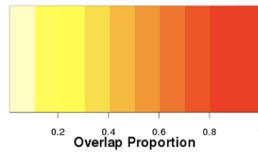
GeneSetDB: output

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



Enrichr



Enrichr

Login | Register
9,169,953 lists analyzed
234,849 terms
128 libraries

Analyze What's New? Libraries Find a Gene About Help

Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum.

Try an example [BED file](#).

No file selected.

Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try two examples:
[crisp set example](#), [fuzzy set example](#)

```
IRF1
CD3Z
CD8A
TBX21
TNFRSF10A
B7H3
CD4
IL10
TGFB1
VEGF
```

18 gene(s) entered

Enter a brief description for the list in case you want to share it. (Optional)

Contribute

Please acknowledge Enrichr in your publications by citing the following references:
Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. 2013;128(14).

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*. 2016; gkw377.

Enrichr

 Enrichr

Login | Register

Transcription Pathways Ontologies Disease/Drugs Cell Types Misc Legacy Crowd

Description No description available (18 genes)

KEGG 2016

Inflammatory bowel disease (IBD)_Homo sapiens_hsa05140
Leishmaniasis_Homo sapiens_hsa05140
T cell receptor signaling pathway_Homo sapiens_hsa05140
Allograft rejection_Homo sapiens_hsa05330
Malaria_Homo sapiens_hsa05144

WikiPathways 2016

Cytokines and Inflammatory Response_Homo sapiens_WP2328
Cytokines and Inflammatory Response (BioCarta)_WP254
Allograft Rejection_Homo sapiens_WP2328
Apoptosis_Mus musculus_WP1254
Apoptosis_Homo sapiens_WP254

ARCHS4 Kinases Coexp

PLK3_human_kinase_ARCHS4_coexpression
PIM3_human_kinase_ARCHS4_coexpression
LCK_human_kinase_ARCHS4_coexpression
MAP3K8_human_kinase_ARCHS4_coexpression
PRKCH_human_kinase_ARCHS4_coexpression

Reactome 2016

TP53 Regulates Transcription of Cell Death C..._Homo sapiens_R-HSA-109581
Extracellular matrix organization_Homo sapiens_hsa04010
Interferon gamma_signaling_Homo sapiens_hsa04010
Apoptosis_Homo sapiens_R-HSA-109581
Programmed Cell Death_Homo sapiens_R-HSA-109581

BioCarta 2016

IFN gamma signaling pathway_Homo sapiens_WP254
Granzyme A mediated Apoptosis Pathway_Homo sapiens_WP254
Apoptotic DNA fragmentation and tissue homoeostasis_WP254
NO2-dependent IL 12 Pathway in NK cells_Homo sapiens_WP254
IL-10 Anti-inflammatory Signaling Pathway_Homo sapiens_WP254

Humancyc 2016

C20 prostanoid biosynthesis_Homo sapiens_WP254

NCI-Nature 2016

IL12 signaling mediated by STAT4_Homo sapiens_WP254
IL12-mediated signaling events_Homo sapiens_WP254
Calcineurin-regulated NFAT-dependent trans..._Homo sapiens_WP254
AP-1 transcription factor network_Homo sapiens_WP254
IL27-mediated signaling events_Homo sapiens_WP254

Panther 2016

Apoptosis signaling pathway_Homo sapiens_WP254
CCKR signaling map ST_Homo sapiens_WP069
Inflammation mediated by chemokine and cytokine_WP254
Interferon-gamma signaling pathway_Homo sapiens_WP254
Toll receptor signaling pathway_Homo sapiens_WP254

BioPlex 2017

CD320
ALDH3B1
MED4
MED14
CNOT2

[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](#)

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

[Previous](#) [Next](#)

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 significantly 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 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 appearing to be enriched in the analysis.