

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?

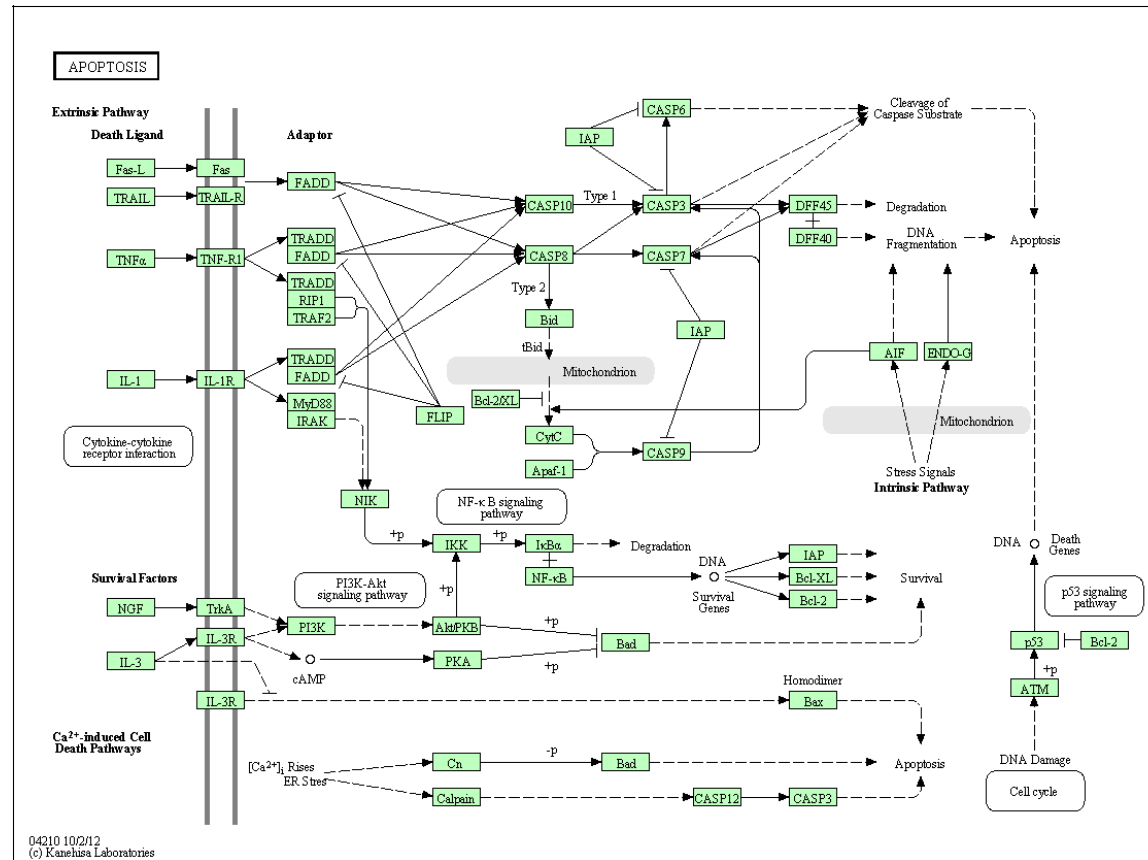
Biological pathways

- 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

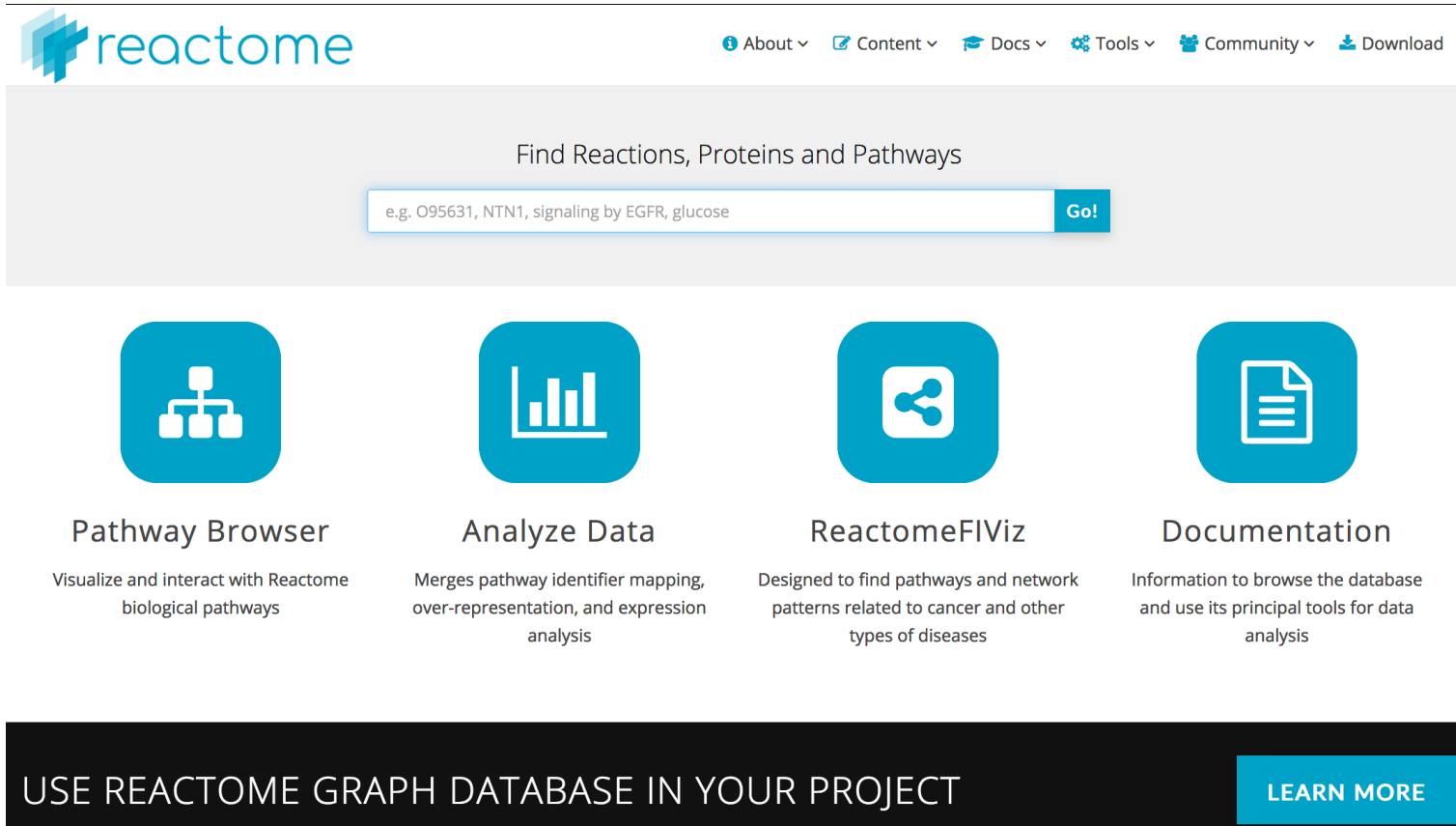
- Kyoto Encyclopedia of Gene and Genomes:
<http://www.genome.jp/kegg/kegg4.html>
- Provides nice (user-created) pathway diagrams (although this is an old database, so the style of the diagrams looks a bit dated).
- XML output includes information about genes involved in pathways, and inter-gene (and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



The image is a screenshot of the Reactome website. At the top left is the Reactome logo, which consists of a blue square with a white 'R' and the word 'reactome' in blue. To the right of the logo is a navigation bar with links: 'About', 'Content', 'Docs', 'Tools', 'Community', and 'Download'. Below the navigation bar is a search bar with the text 'Find Reactions, Proteins and Pathways'. The search bar contains the example text 'e.g. O95631, NTN1, signaling by EGFR, glucose' and a 'Go!' button. Below the search bar are four icons in blue squares: a tree diagram for 'Pathway Browser', a bar chart for 'Analyze Data', a network diagram for 'ReactomeFIViz', and a document icon for 'Documentation'. Each icon has a title and a description below it. At the bottom of the page is a black banner with the text 'USE REACTOME GRAPH DATABASE IN YOUR PROJECT' and 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

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

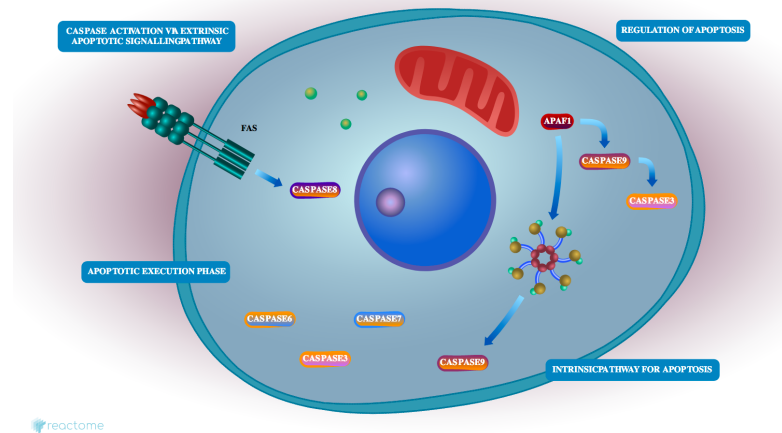
Documentation
Information to browse the database and use its principal tools for data analysis

USE REACTOME GRAPH DATABASE IN YOUR PROJECT

LEARN MORE

<http://www.reactome.org>

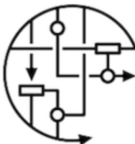
Reactome: apoptosis pathway



- 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



search

search

- Help
- About us
- Contact us
- Report a bug
- How to cite

download

- Download files
- Web service API
- WikiPathways RDF
- Embed code

activity

- Browse pathways
- Recent changes
- New pathways
- Edit pathways
- Create pathway
- Statistics

tools

- PathwayWidget
- Pathway Finder
- Software tools

community

- Quality control
- Development
- WikiPathways Blog
- AOP portal
- CIRM portal
- COVID-19

page | discussion | view source | history

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

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

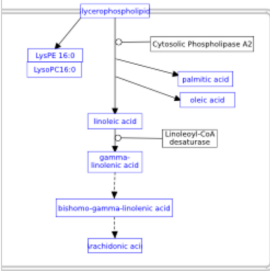
Multiple formats and methods

Growth

New pathways added each month

Today's Featured Pathway

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




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

The diagram illustrates the metabolic pathway of linoleic acid. It starts with LysPE 16:0 and LysPC 16:0, which are converted to LysPE 18:0 and LysPC 18:0 by the enzyme Cytosolic Phospholipase A2. These are then converted to linoleic acid, which is further processed to gamma-linolenic acid and then to arachidonic acid. The pathway is labeled as 'Linoleic acid metabolism known to be affected by coronavirus infection (Homo sapiens)'.

Linoleic acid metabolism known to be affected by coronavirus infection

Curator of the Week

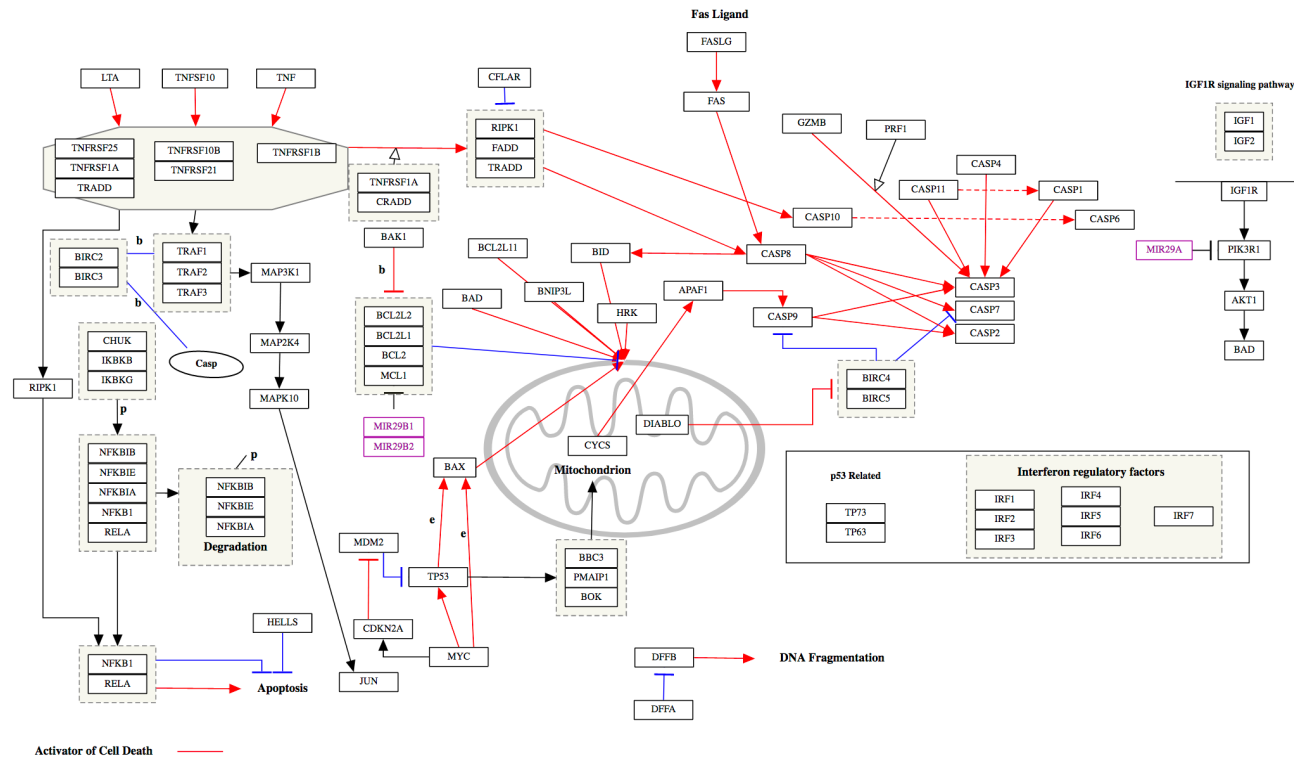


Elisson Lopes (UFMG, Brazil)

<http://www.wikipathways.org>

9/27

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

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](#).

Browse...

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 .

Submit



KEGG 2016

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

WikiPathways 2016

Cytokines and Inflammatory Response_Hom
Cytokines and Inflammatory Response (BioC
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
MAP3K3_human_kinase_ARCHS4_coexpressi
PRKCH_human_kinase_ARCHS4_coexpressio

Reactome 2016

TP53 Regulates Transcription of Cell Death C
Extracellular matrix organization_Homo sap
Interferon gamma signaling_Homo sapiens_
Apoptosis_Homo sapiens_R-HSA-109581
Programmed Cell Death_Homo sapiens_R-H

BioCarta 2016

IFN gamma signaling pathway_Homo sapie
Granzyme A mediated Apoptosis Pathway_H
Apoptotic DNA fragmentation and tissue ho
NO2-dependent IL 12 Pathway in NK cells_H
IL-10 Anti-Inflammatory Signaling Pathway_I

Humancyc 2016

C20 prostanoil biosynthesis_Homo sapiens

NCI-Nature 2016


IL12 signaling mediated by STAT4_Homo sap
IL12-mediated signaling events_Homo sapie
Calcineurin-regulated NFAT-dependent tran
AP-1 transcription factor network_Homo sap
IL27-mediated signaling events_Homo sapie

Panther 2016

Apoptosis signaling pathway_Homo sapiens
CCKR signaling map ST_Homo sapiens_P069
Inflammation mediated by chemokine and c
Interferon-gamma signaling pathway_Homo
Toll receptor signaling pathway_Homo sapie

BioPlex 2017

CD320
ALDH3B1
MED4
MED14
CNOT2


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

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 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.