The Bio-Data Club @ Moffitt

Sponsored by the R consortium

WHO WE ARE: Hosted by the Moffitt Biostatistics and Bioinformatics Shared Resource, the club is an informal group at Moffitt dedicated to promoting a fun, supportive, and hands-on environment for learning biological data science. Our meetings are intended for all levels of skill and computational backgrounds.

March 2019 meeting topic:
Automate your pathway enrichment analysis with R

Date/time: Friday, March 15th @ 2pm

Location: Moffitt Stabile Research Building (SRB), David Murphey Conference Room (1st floor)

We meet once a month on Fridays. Bring a laptop! For a complete schedule and contact information, please visit the club GitHub page at https://pstew.github.io/biodataclub

PREVIOUS MEETINGS HAVE FEATURED...

Reproducible research with git and GitHub

Building interactive web applications with R/Shiny

Introduction to t-SNE clustering

MOFFITT CANCER CENTER

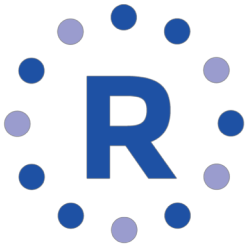
Automate your pathway enrichment analysis with R

2019-03-14

Paul Stewart, PhD

Bioinformatics Staff Scientist, BBSR

Sponsored by



consortium

Why pathway analysis?

- Omics experiments can result in large lists of genes/proteins
  - Immediate relevance unclear
  - Relationship to one another unclear
- Genes can be grouped according to biological function, cellular compartment, chromosome, etc.
- Pathways help condense gene lists and prevent information overload.

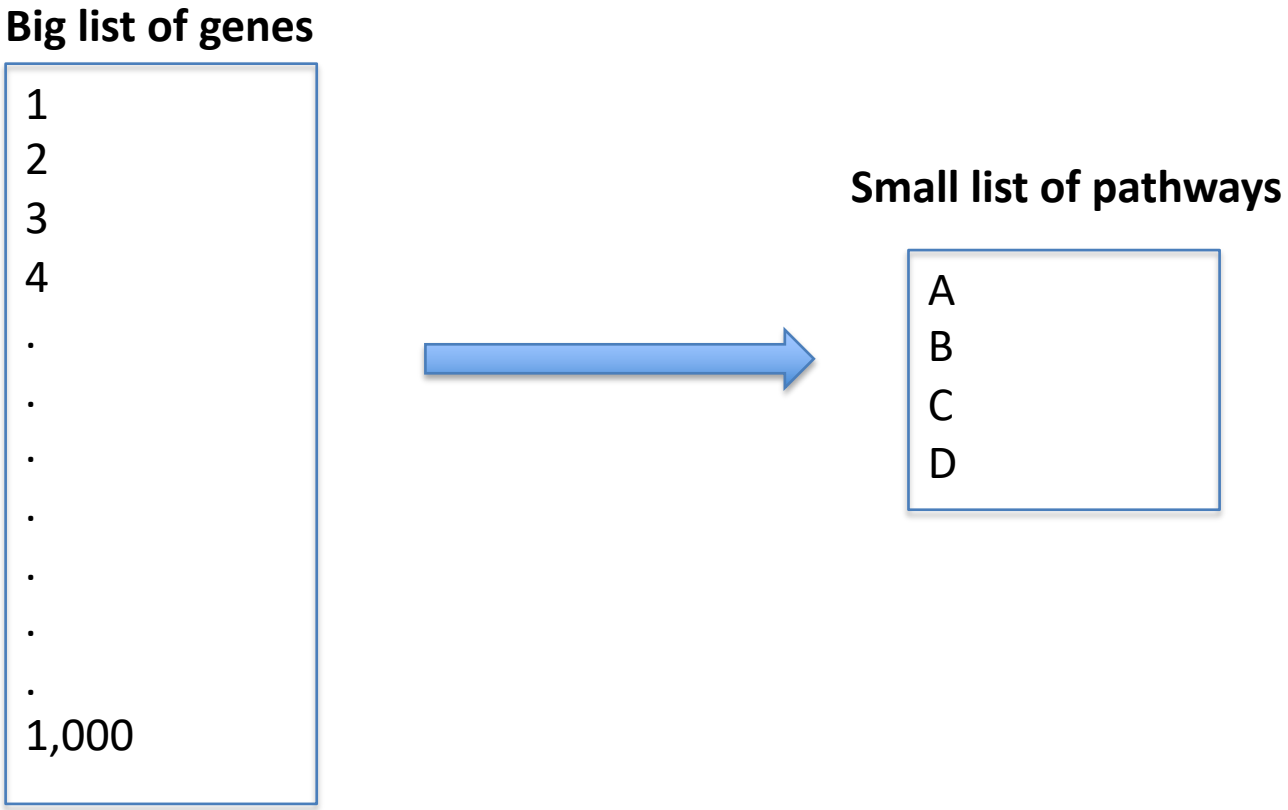
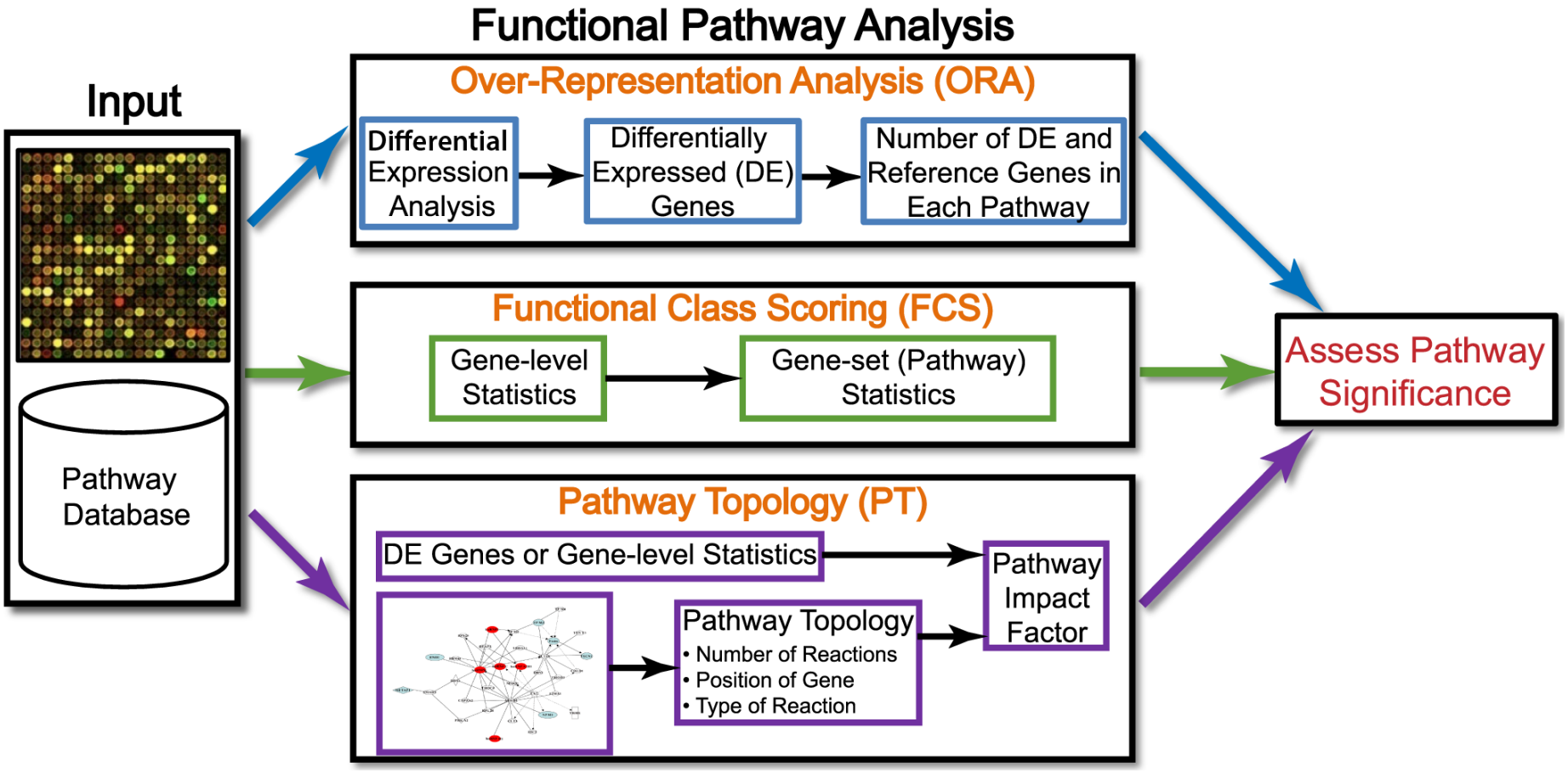


Figure 1. Overview of existing pathway analysis methods using gene expression data as an example.



Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLOS Computational Biology 8(2): e1002375. https://doi.org/10.1371/journal.pcbi.1002375
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002375

Over Representation Analysis

- In an experiment, you measured the expression of 30 genes.
- 15 differentially expressed genes were identified,
- 15 genes were associated with the GO pathway ‘‘DNA-templated transcription, elongation”.
- 12 genes are both differentially expressed and tagged with “translation initiation”.
  - Is this statistically significant?

Contingency table for gene expression data

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	12	3	15
NOT IN Transcription Elongation	3	12	15
Total	15	15	30

http://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/

Over Representation Analysis Math (Fisher's Exact Test)

Contingency table for gene expression data

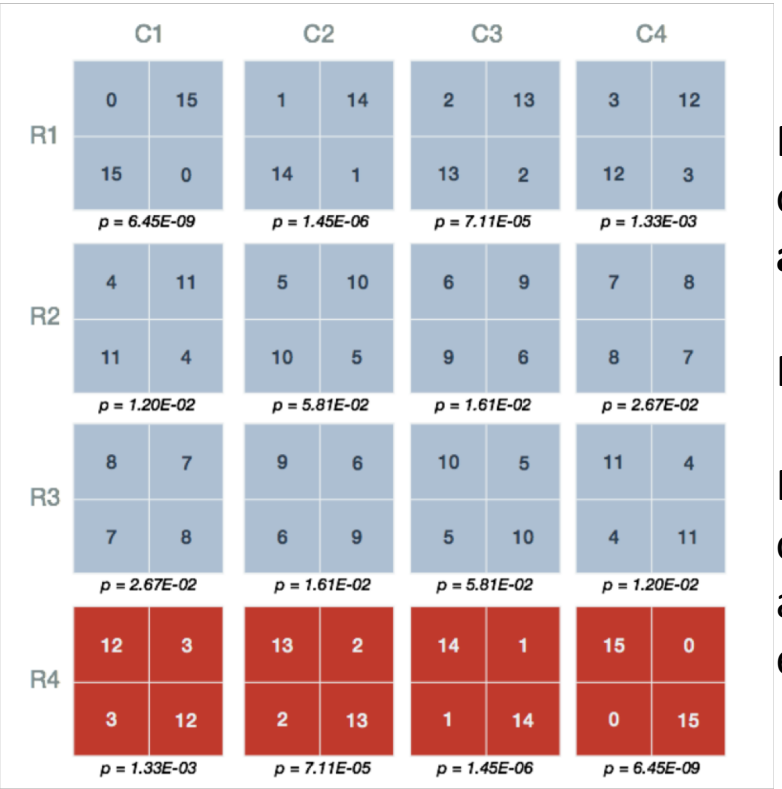
	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	12	3	15
NOT IN Transcription Elongation	3	12	15
Total	15	15	30

"n choose k" = $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

What is the probability for the table to the left?

$$P = \frac{(15 \text{ choose } 12) * (15 \text{ choose } 3)}{(30 \text{ choose } 15)}$$

$$= 1.33\text{E-}03$$



Fisher's Exact Test amounts to summing the probability of observing our table of observed joint values **in addition to those more extreme than our table, so**

P-value = (R4, C1) + (R4, C2) + (R4, C3) + (R4, C4) = 0.0014

From this result, we claim that the probability of our observed data or that more extreme under the assumption that there is no association between expression and gene set membership is 0.0014.

http://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/

Enrichr

Login | Register

8,579,126 lists analyzed
232,994 terms
125 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.

Choose File | No file chosen

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

TALDO1
ALDH3A1
DDA1
GSTA2
PCYT1A
GPC1
GSTM3
GSTM5
PFN2
PFN2

93 gene(s) entered

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

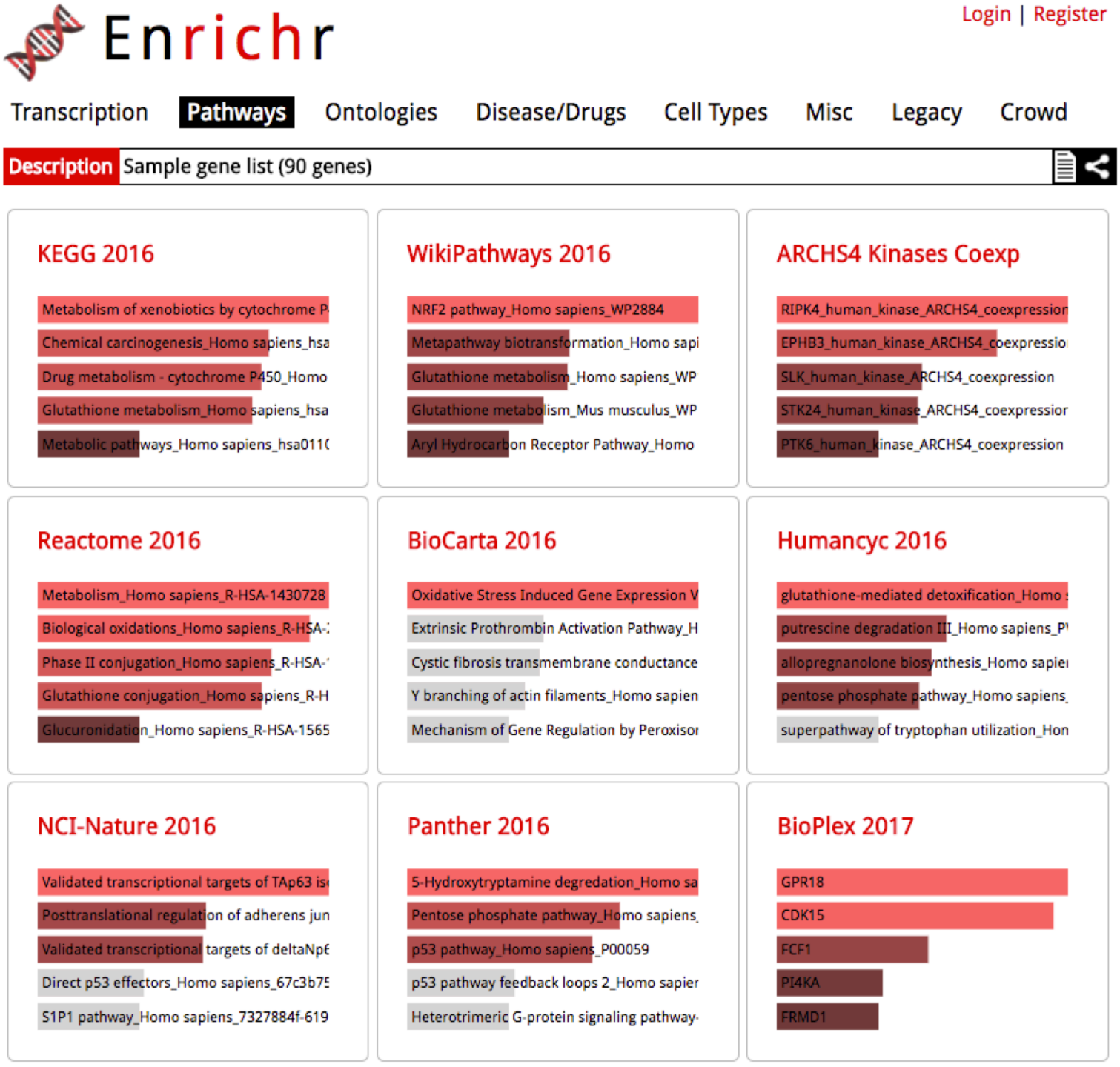
Submit

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* .

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


Enrichr
Login | Register
Transcription
Pathways
Ontologies
Disease/Drugs
Cell Types
Misc
Legacy
Crowd
Description
Sample gene list (90 genes)
KEGG 2016
Metabolism of xenobiotics by cytochrome P
Chemical carcinogenesis_Homo sapiens_hsa
Drug metabolism - cytochrome P450_Homo
Glutathione metabolism_Homo sapiens_hsa
Metabolic pathways_Homo sapiens_hsa011(
WikiPathways 2016
NRF2 pathway_Homo sapiens_WP2884
Metapathway biotransformation_Homo sapi
Glutathione metabolism_Homo sapiens_WP
Glutathione metabolism_Mus musculus_WP
Aryl Hydrocarbon Receptor Pathway_Homo
ARCHS4 Kinases Coexp
RIPK4_human_kinase_ARCHS4_coexpression
EPHB3_human_kinase_ARCHS4_coexpressio
SLK_human_kinase_ARCHS4_coexpression
STK24_human_kinase_ARCHS4_coexpressior
PTK6_human_kinase_ARCHS4_coexpression
Reactome 2016
Metabolism_Homo sapiens_R-HSA-1430728
Biological oxidations_Homo sapiens_R-HSA-
Phase II conjugation_Homo sapiens_R-HSA-
Glutathione conjugation_Homo sapiens_R-H
Glucuronidation_Homo sapiens_R-HSA-1565
BioCarta 2016
Oxidative Stress Induced Gene Expression V
Extrinsic Prothrombin Activation Pathway_H
Cystic fibrosis transmembrane conductance
Y branching of actin filaments_Homo sapien
Mechanism of Gene Regulation by Peroxisor
Humancyc 2016
glutathione-mediated detoxification_Homo
putrescine degradation III_Homo sapiens_P
allopregnanolone biosynthesis_Homo sapie
pentose phosphate pathway_Homo sapiens_
superpathway of tryptophan utilization_Hon
NCI-Nature 2016
Validated transcriptional targets of TAp63 is
Posttranslational regulation of adherens jun
Validated transcriptional targets of deltaNp6
Direct p53 effectors_Homo sapiens_67c3b75
S1P1 pathway_Homo sapiens_7327884f-619
Panther 2016
5-Hydroxytryptamine degredation_Homo sa
Pentose phosphate pathway_Homo sapiens_
p53 pathway_Homo sapiens_P00059
p53 pathway feedback loops 2_Homo sapier
Heterotrimeric G-protein signaling pathway-
BioPlex 2017
GPR18
CDK15
FCF1
PI4KA
FRMD1

Description Sample gene list (90 genes)

GO Cellular Component 2017b

type III intermediate filament (GO:0045098)
glial filament (GO:0097426)
lamin filament (GO:0005638)
neurofilament (GO:0005883)
keratin filament (GO:0045095)

GO Biological Process 2017b

glutathione biosynthetic process (GO:00067
epidermal cell differentiation (GO:0009913)
glutathione derivative biosynthetic process (
trypanothione biosynthetic process (GO:001
glutathione deglycation (GO:0036531)

GO Molecular Function 2017b

glutathione transferase activity (GO:000436
protein homodimerization activity (GO:0042
NAD(P)H:methyl-1,4-benzoquinone oxidorec
p-benzoquinone reductase (NADPH) activity
NADPH:quinone reductase activity (GO:0003

MGI Mammalian Phenotype 2017

MP:0001236_abnormal_epidermis_stratum_
MP:0001730_embryonic_growth_arrest
MP:0000764_abnormal_tongue_epithelium_
MP:0003809_abnormal_hair_shaft_morphol
MP:0000352_decreased_cell_proliferation

Human Phenotype Ontology

Palmoplantar keratoderma (HP:0000982)
Anonychia (HP:0001798)
Abnormal blistering of the skin (HP:0008066
Woolly hair (HP:0002224)
Skin ulcer (HP:0200042)

Jensen TISSUES

Gut
BTO:0004410
Mouth
Blood_platelet
Bronchial_epithelial_cell

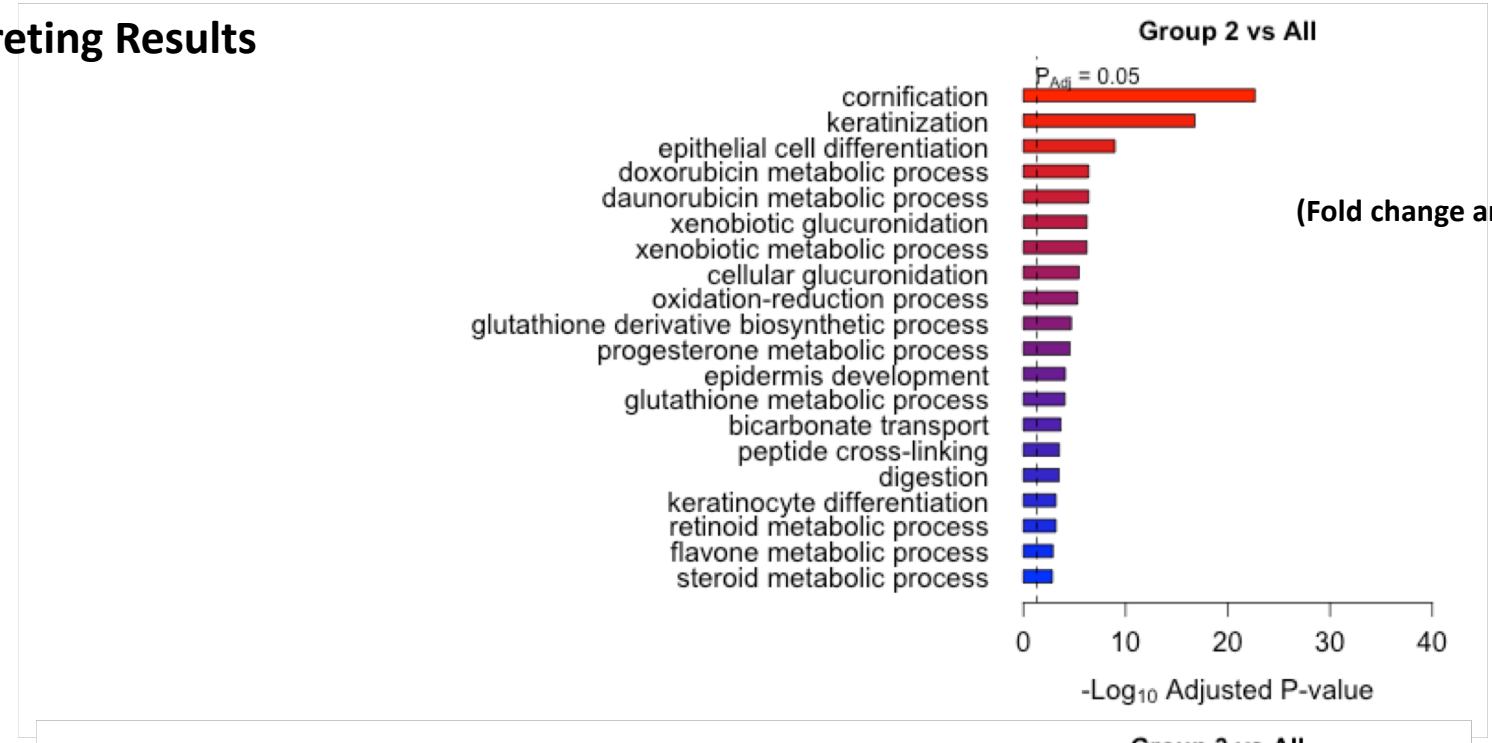
Jensen COMPARTMENTS

Extracellular_exosome
Extracellular_vesicle
Extracellular_organelle
Extracellular_region_part
Extracellular_region

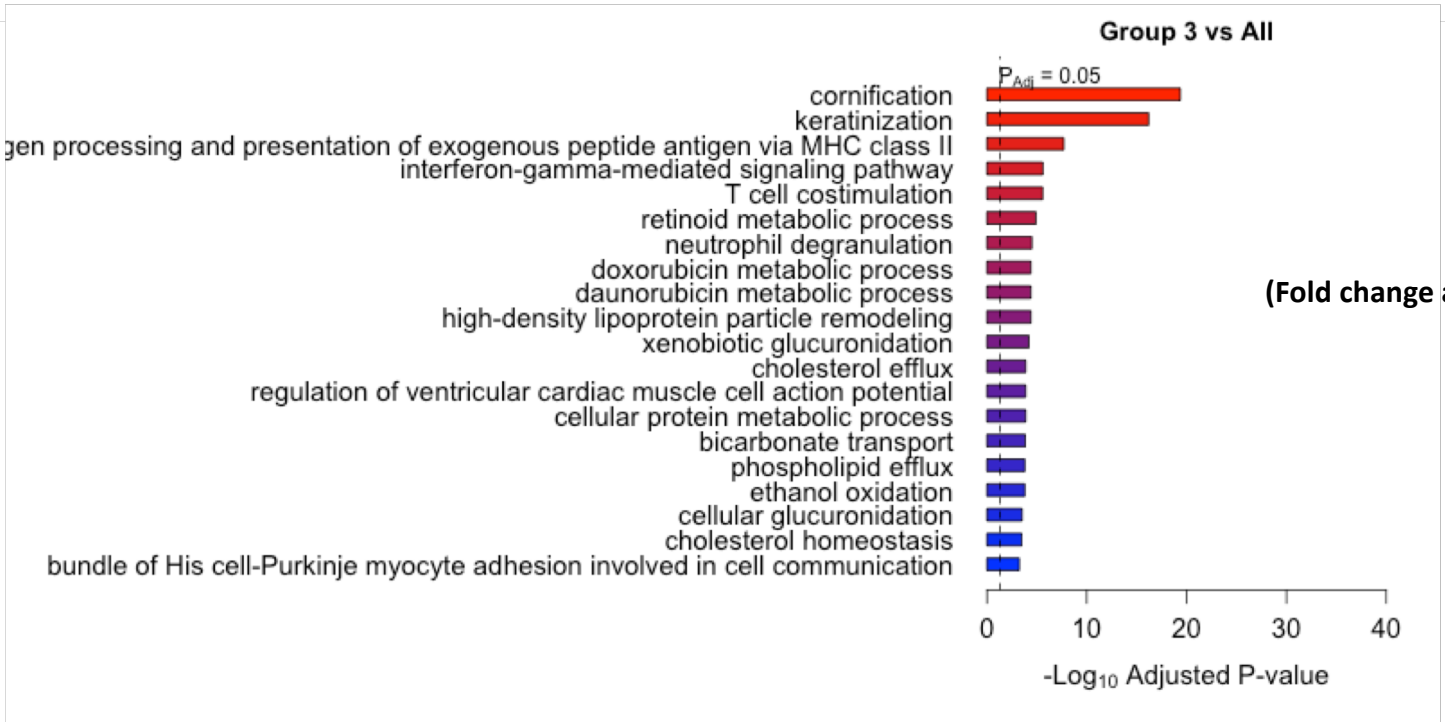
Jensen DISEASES

Palmoplantar_keratosis
Anemia
Pericholangitis
Bilirubin_metabolic_disorder
Fissured_tongue

Interpreting Results

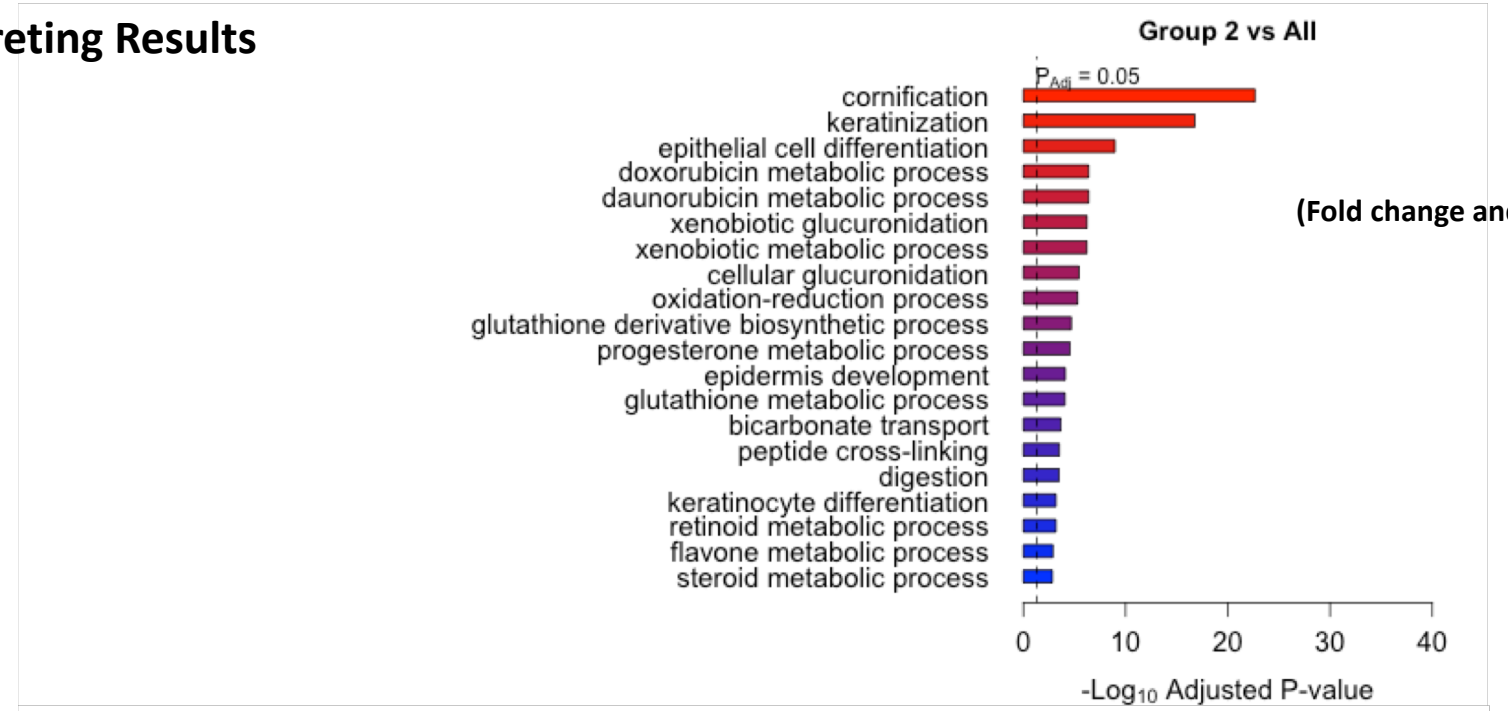


(Fold change and p-value cutoff)

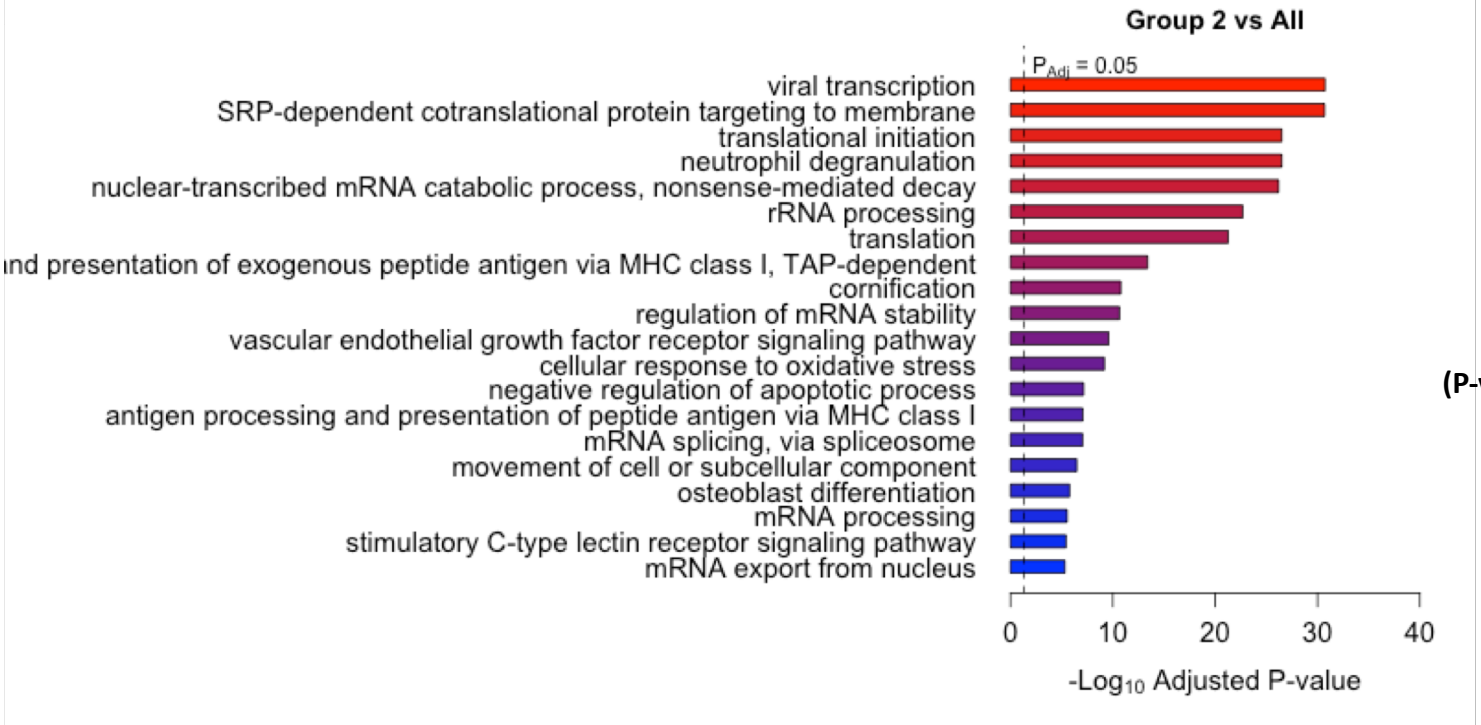


(Fold change and p-value cutoff)

Interpreting Results



(Fold change and p-value cutoff)



(P-value cutoff only)

Setting cutoffs

Dalman *et al. BMC Bioinformatics* 2012, **13**(Suppl 2):S11
http://www.biomedcentral.com/1471-2105/13/S2/S11

BMC Bioinformatics

PROCEEDINGS Open Access

Fold change and p-value cutoffs significantly alter microarray interpretations

Mark R Dalman[1*], Anthony Deeter[2], Gayathri Nimishakavi[2], Zhong-Hui Duan[2]

From Great Lakes Bioinformatics Conference 2011
Athens, OH, USA. 2-4 May 2011

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2654802/

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 6 2009, pages 765–771
doi:10.1093/bioinformatics/btp053

Gene expression

Testing significance relative to a fold-change threshold is a TREAT

Davis J. McCarthy and Gordon K. Smyth*
The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3050, Australia

Received on October 12, 2008; revised on January 21, 2009; accepted on January 22, 2009
Advance Access publication January 28, 2009
Associate Editor: Joaquin Dopazo

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3305783/