

Extracting Biological Information from Gene Lists

Simon Andrews, Laura Biggins, Boo Virk

simon.andrews@babraham.ac.uk

laura.biggins@babraham.ac.uk

v2025-07

Programme

- The theory and practice of gene set enrichment
- Gene set enrichment practical
- Presenting results
- Dealing with artefacts and biases
- Motif analysis
- Motif analysis practical

Standard Gene List Output

Rank	Well	Gene Name	P-value	GPR Fold Change	GPR Fold Change Graph	Control 1	Control 2	Control 3
1	E10	<u>Tnfrsf18</u>	0.006689	-7.708526		26.291138	25.415058	25.808804
2	E08	<u>Ly9</u>	0.009059	-7.238955		25.672344	24.660522	24.845451
3	E01	<u>Tollip</u>	0.081636	-14.769324		27.33491	31.586285	27.811256
4	H10	<u>Stat3</u>	0.092269	-2.377623		25.84287	26.284285	26.874344
5	F03	<u>Nt5e</u>	0.097510	-1.511391		25.420982	26.977015	25.08718
6	C01	<u>Tnfrsf1b</u>	0.099746	-4.026225		40.0	37.44099	36.49696
7	A09	<u>Ccnd3</u>	0.100523	3.755167		30.837646	30.475822	30.468536
8	D11	<u>Nfatc2</u>	0.124354	5.534758		28.610485	29.669998	30.464863
9	D05	<u>Il2ra</u>	0.132781	-1.549923		37.23	35.44099	35.49696
10	H04	<u>Sema4a</u>	0.133853	-5.447223		36.48277	36.928036	32.373432
11	D01	<u>Tnfsf4</u>	0.144796	6.022623		21.888157	20.845629	22.976254
12	B10	<u>Nfat5</u>	0.145780	8.067699		30.449022	30.795446	30.850525
13	D07	<u>Cd3e</u>	0.166966	5.300400		28.893595	28.981432	30.581322
14	H05	<u>Nrp1</u>	0.171774	3.802116		30.856043	30.58041	30.099209
15	G05	<u>Cd53</u>	0.180716	-2.249306		33.33491	33.586285	33.811256
16	D09	<u>Cd28</u>	0.188418	-4.313547		24.510563	23.23	20.464325
17	D02	<u>Pou2af1</u>	0.199099	2.734895		26.449022	22.795446	23.850525
18	F09	<u>Gadd45b</u>	0.209415	-1.859485		25.837646	25.475822	24.468536
19	D04	<u>S100a6</u>	0.221836	-1.869103		22.482086	24.83037	23.917696
20	B07	<u>Stat6</u>	0.233153	-1.493636		33.44925	32.16483	32.71563

Descriptions aren't always informative

Gene	Description
Gpr55	G protein-coupled receptor 55 [Source:MGI Symbol;Acc:MGI:2685064]
Ncl	nucleolin [Source:MGI Symbol;Acc:MGI:97286]
Aspm	asp (abnormal spindle)-like, microcephaly associated (Drosophila) [Source:MGI Symbol;Acc:MGI:1334448]
Tnfsf4	tumor necrosis factor (ligand) superfamily, member 4 [Source:MGI Symbol;Acc:MGI:104511]
Ephx1	epoxide hydrolase 1, microsomal [Source:MGI Symbol;Acc:MGI:95405]
Setx	senataxin [Source:MGI Symbol;Acc:MGI:2443480]
Angptl2	angiopoietin-like 2 [Source:MGI Symbol;Acc:MGI:1347002]
Ggta1	glycoprotein galactosyltransferase alpha 1, 3 [Source:MGI Symbol;Acc:MGI:95704]
Dab2ip	disabled homolog 2 (Drosophila) interacting protein [Source:MGI Symbol;Acc:MGI:1916851]
Neb	nebulin [Source:MGI Symbol;Acc:MGI:97292]
Ernm	ermin, ERM-like protein [Source:MGI Symbol;Acc:MGI:1925017]
Ckap5	cytoskeleton associated protein 5 [Source:MGI Symbol;Acc:MGI:1923036]
Prr5l	proline rich 5 like [Source:MGI Symbol;Acc:MGI:1919696]
Arhgap11a	Rho GTPase activating protein 11A [Source:MGI Symbol;Acc:MGI:2444300]
Bub1b	budding uninhibited by benzimidazoles 1 homolog, beta (S. cerevisiae) [Source:MGI Symbol;Acc:MGI:1333889]
Prnp	prion protein [Source:MGI Symbol;Acc:MGI:97769]
Fam102b	family with sequence similarity 102, member B [Source:MGI Symbol;Acc:MGI:3036259]

Gene summary sites are useful for single genes

TNFSF4 - tumor necrosis factor (ligand) superfamily...

Homo sapiens

Synonyms: CD134L, CD252, GP34, ...

Glycoprotein Gp34, OX-40L, ...

[Biagi, E. et al.](#), [Godfrey, W.R. et al.](#), [Wang, X. et al.](#), [Takasawa, N. et al.](#), [Ito, T. et al.](#), [et al.](#)

Welcome! If you are familiar with the subject of this article, you can contribute to this open access knowledge base by deleting incorrect information, restructuring or completely rewriting any text. [Read more.](#)

Disease relevance of TNFSF4

- In two independent human populations, the less common allele of SNP rs3850641 in TNFSF4 was significantly more frequent ($P \leq 0.05$) in individuals with [myocardial infarction](#) than in controls [1].
- However, [cytotoxic T lymphocyte](#) (CTL) clones specific for [Epstein-Barr virus](#) (EBV)-transformed autologous lymphoblastic [cell lines](#) (LCLs) induced both OX40 and OX40L expression after antigen or [T cell](#) receptor (TCR) stimulation [2].
- We have cloned and sequenced a cDNA encoding gp34, a novel [glycoprotein](#) expressed in cells bearing human [T-cell](#) leukemia virus type I (HTLV-I) [3].
- On the other hand, gp34 was not expressed on these cells, although its expression is also known to be associated with [HTLV-I-infection](#) [4].
- Regulation of [T cell](#) activation [in vitro](#) and in vivo by targeting the OX40-OX40 ligand interaction: amelioration of ongoing [inflammatory bowel disease](#) with an OX40-IgG fusion protein, but not with an OX40 ligand-IgG fusion protein [5].

High impact information on TNFSF4

- We therefore conclude that Tnfsf4 underlies [Ath1](#) in mice and that polymorphisms in its human homolog TNFSF4 increase the risk of [myocardial infarction](#) in humans [1].
- The [quantitative trait locus](#) region encompasses 11 known genes, including Tnfsf4 (also called [Ox40](#) or Cd134l), which encodes OX40 ligand [1].
- When activated in the presence of leukemic CLL [B cells](#), [T cells](#) rapidly up-regulate CD30 through an [OX40 ligand](#) and [interleukin 4](#) (IL-4)-dependent mechanism [6].
- Here we report that [TSLP induced](#) human DCs to express [OX40 ligand](#) (OX40L) but not IL-12 [7].
- [TSLP-induced](#) OX40L on DCs was required for [triggering](#) naive CD4(+) [T cells](#) to [produce](#) [IL-4](#), -5, and -13 [7].

Biological context of TNFSF4

- This study suggests a possible function of OX40L / OX40, through T cell-T cell interaction, in the reactivation of memory [T cells](#) in an autocrine manner, with implications for the pathogenesis of viral infections and neoplasms [2].
- These results indicated that rat OX40L can provide an efficient costimulation for rat [T cells](#) and that it may be involved in HTLV-1-associated [pathologies](#) in the rat system as has been suggested in the human system [8].
- Enhancing the immunostimulatory function of [dendritic cells](#) by [transfection](#) with mRNA encoding OX40 ligand [9].
- [Cell adhesion](#) assay was performed and in at least three cases, fresh ATL cells exhibited adhesion to human [umbilical vein endothelial cells](#) that could be considerably inhibited by either anti-OX40 MoAb or anti-gp34 MoAb [4].
- [T cell](#) proliferation by direct cross-talk between [OX40 ligand](#) on human [mast cells](#) and [OX40](#) on human [T cells](#): comparison of [gene expression](#) profiles between human tonsillar and lung-cultured [mast cells](#) [10].



LLMs can be useful if used with caution

What is known about the functionality of human TNFSF4?

Human **TNFSF4**, also known as **OX40 ligand (OX40L)** or **CD252**, is a member of the tumor necrosis factor (TNF) superfamily. It is a costimulatory molecule that plays a critical role in the immune system by influencing T-cell activation, survival, and differentiation. Here are the key details about its functionality:

Structure and Expression

- TNFSF4 is expressed as a type II transmembrane protein on activated antigen-presenting cells (APCs), such as dendritic cells, B cells, and macrophages.
- It interacts with its receptor, **OX40 (CD134)**, which is expressed on activated T cells, particularly CD4+ T cells.

1. T-cell Costimulation

- TNFSF4 acts as a costimulatory signal during T-cell activation, enhancing the proliferation, survival, and function of T cells.
- It amplifies the immune response by promoting cytokine production, such as IL-2, which is crucial for T-cell growth.

2. T-cell Memory Development

- TNFSF4-OX40 signaling supports the development and maintenance of memory T cells, ensuring a rapid and robust immune response upon re-exposure to the same antigen.

3. Effector T-cell Function

- This pathway enhances the function of effector T cells, including their ability to produce cytokines like IFN- γ and TNF- α , which are essential for combating infections and tumors.

4. Regulatory T-cell (Treg) Modulation

- TNFSF4 can influence Treg function, sometimes promoting their suppressive activity to prevent excessive immune responses.

5. Role in Immune Diseases

- Dysregulated TNFSF4 signaling is implicated in autoimmune diseases (e.g., systemic lupus erythematosus, rheumatoid arthritis) due to its role in sustaining chronic immune activation.
- It may also contribute to allergic diseases and inflammation.

6. Role in Cancer Immunity

- TNFSF4 expression on APCs can enhance anti-tumor immunity by supporting T-cell-mediated immune responses against tumors.



ChatGPT

Functional analysis relates hits to existing knowledge

Germ-line stem cell division

The self-renewing division of a germline stem cell to produce a daughter stem cell and a daughter germ cell, which will divide to form the gametes.

Gene/product	Gene/product name
Hoxc4	homeobox C4
Ing2	inhibitor of growth family, member 2
Stra8	stimulated by retinoic acid gene 8
Zbtb16	zinc finger and BTB domain containing 16
Etv5	ets variant 5

Advantages:

- Biological insight
- Validation of experiment
- Generate new hypotheses

Limitations:

- You can only discover what is already known
 - Novel functionality will be missing
 - Existing annotations may be incorrect
 - Many species are poorly supported

Functionality is generally annotated on genes

- Things to think about
 - Converting hits to genes
 - Transcripts / Proteins are easy
 - Genomic positions may be possible
 - Gene nomenclature
 - Names change over time
 - Gene definitions appear / change

- Types of list
 - Categorical (hit or not a hit)
 - Ordered
 - Quantitative

Hits

ABC1
DEF1
GHI1
JKL1

[All non hits]

Ordered

1. DEF1
2. ABC1
3. JKL1
4. GHI1

[All non hits]

Quant

ABC1 = 5.3
DEF1 = 2.1
GHI1 = 7.9
JKL1 = 1.0
MNO1 = 0.4
PQR1 = 5.7
STU1 = 3.8

Comparing your hits to functional gene sets

Germ-line stem cell division

The self-renewing division of a germline stem cell to produce a daughter stem cell and a daughter germ cell, which will divide to form the gametes.

Gene/product	Gene/product name
--------------	-------------------

Hoxc4	homeobox C4
Ing2	inhibitor of growth family, member 2
Stra8	stimulated by retinoic acid gene 8
Zbtb16	zinc finger and BTB domain containing 16
Etv5	ets variant 5

A4galt

Atl1

Cdk19

Cdon

Cecr2

Etv5

My Hits

Flywch1

Gnpda2

Hoxc4

Ing2

ligp1

Map3k9

Mypop

Rnf6

Serinc1

Stra8

Trp73

Zbtb16

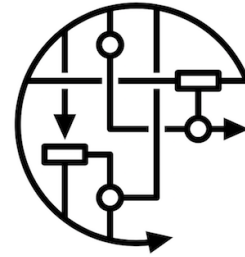
Nothing is ever straight forward...

Best hit: “DNA Methylation” $p < 2e-10$

- name: DNA methylation
- datasource: reactome
- organism: Human
- idtype: hgnc symbol
- Genes:
- Methyltransferases: DNMT1 DNMT3A DNMT3B DNMT3L
- Methyltransferase targeting protein: UHRF1
- **Histones!!!** H2AFB1 H2AFJ H2AFV H2AFX H2AFZ H2BFS H3F3A H3F3B HIST1H2AB HIST1H2AC HIST1H2AD HIST1H2AE HIST1H2AJ HIST1H2BA HIST1H2BB HIST1H2BC HIST1H2BD HIST1H2BE HIST1H2BF HIST1H2BG HIST1H2BH HIST1H2BI HIST1H2BJ HIST1H2BK HIST1H2BL HIST1H2BM HIST1H2BN HIST1H2BO HIST1H3A HIST1H3B HIST1H3C HIST1H3D HIST1H3E HIST1H3F HIST1H3G HIST1H3H HIST1H3I HIST1H3J HIST1H4A HIST1H4B HIST1H4C HIST1H4D HIST1H4E HIST1H4F HIST1H4H HIST1H4I HIST1H4J HIST1H4K HIST1H4L HIST2H2AA3 HIST2H2AA4 HIST2H2AC HIST2H2BE HIST2H3A HIST2H3C HIST2H3D HIST2H4A HIST2H4B HIST3H2BB HIST4H4

Sources of functional gene lists

- Human curated
 - Gene Ontology
 - Biological Pathways
- Domains / Patterns
 - Protein functional domains
 - Transcription factor regulated
- Experimental
 - Co-expressed genes
 - Interactions
 - Hits from other studies



WIKIPATHWAYS
Pathways for the People



GENEONTOLOGY
Unifying Biology



InterPro
Protein sequence analysis & classification



STRING



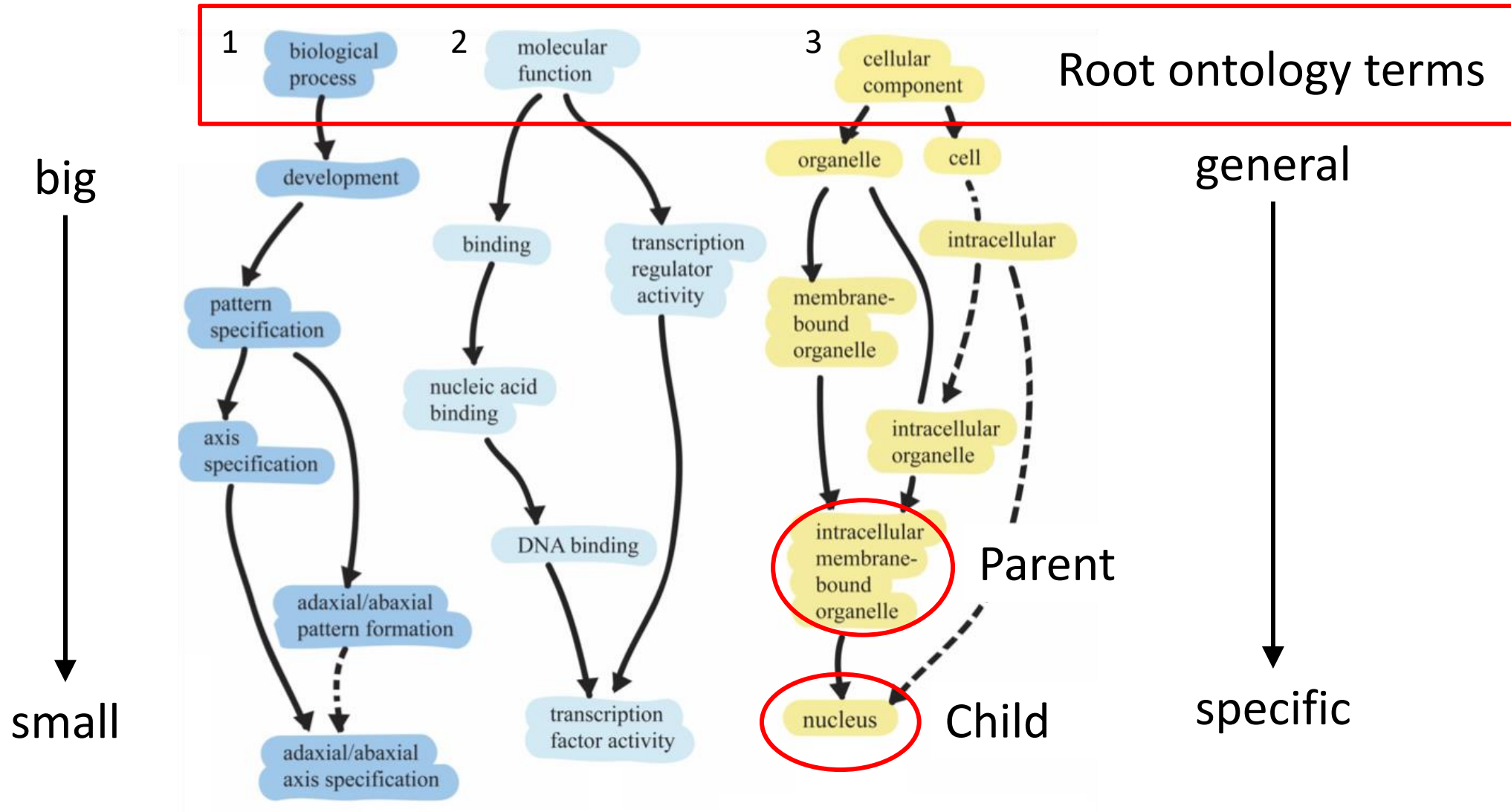
COXPRESdb



Gene Ontology is a human curated functional database



GO has three domains and a hierarchical structure



Genes are specifically placed into each domain

Nanog homeobox

- **Cellular Component**

- GO:0005634 nucleus
- GO:0005654 nucleoplasm
- GO:0005730 nucleolus

- **Molecular Function**

- GO:0003677 DNA binding
- GO:0003700 transcription factor activity
- GO:0003714 transcription corepressor activity
- GO:0005515 protein binding
- GO:0043565 sequence-specific DNA binding

- **Biological Process**

































- GO:0001714 endodermal cell fate specification
- GO:0006351 transcription, DNA-templated
- GO:0006355 regulation of transcription, DNA-templated
- GO:0007275 multicellular organism development
- GO:0008283 cell proliferation
- GO:0019827 stem cell population maintenance
- GO:0030154 cell differentiation
- GO:0035019 somatic stem cell population maintenance
- GO:0045595 regulation of cell differentiation
- GO:0045944 positive regulation of transcription from RNA pol2
- GO:1903507 negative regulation of nucleic acid-templated transcription

GO Annotations come with evidence

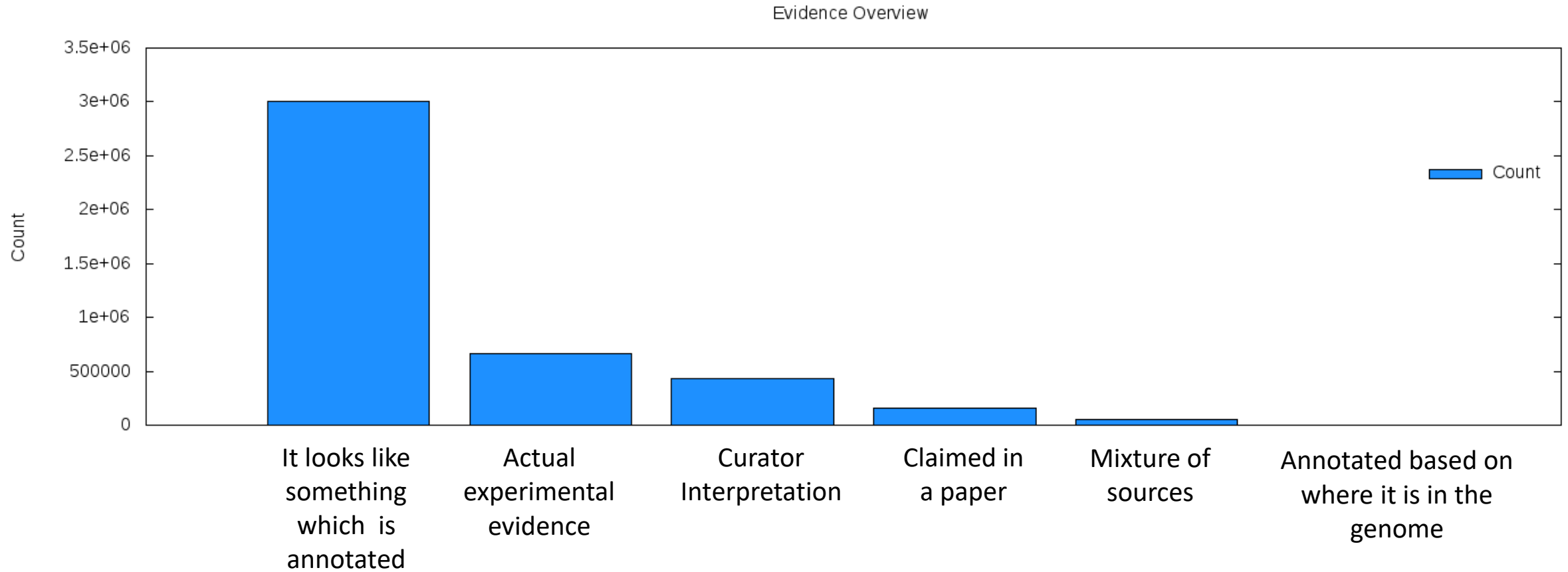
- Experimental
 - Experiment (EXP)
 - Direct Assay (IDA)
 - Physical Interaction (IPI)
 - Mutant Phenotype (IMP)
- Computational
 - Sequence Similarity (ISS)
 - Sequence Model (ISM)
 - Genomic Context (IGC)
 - Biological aspect of Ancestor (IBA)
 - Key Residues (IKR)

- Publications
- Curators

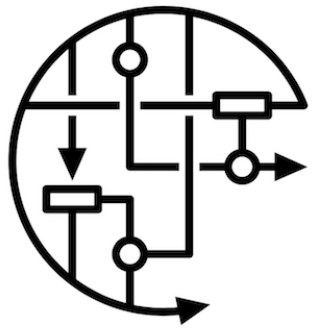
Gene Ontology

					Inferred from experiment [IDA, IPI, IMP, IGI, IEP]
					Direct assay [IDA], Mutant phenotype [IMP]
					Genetic interaction [IGI], Physical interaction [IPI]
					Inferred from High Throughput Experiment [HDA, HMP, HGI, HEP]
					High Throughput Direct Assay [HDA], High Throughput Mutant Phenotype [HMP]
					High Throughput Genetic interaction [HGI], High Throughput Expression pattern [HEP]
					Traceable author [TAS], Non-traceable author [NAS], Inferred by curator [IC]
					Expression pattern [IEP], Sequence or structural similarity [ISS], Genomic context [IGC]
					Sequence Model [ISM], Sequence Alignment [ISA], Sequence Orthology [ISO]
					Biological aspect of ancestor [IBA], Rapid divergence [IRD]
					Reviewed computational analysis [RCA], Electronic annotation [IEA]
					No biological data [ND], Not annotated or not in background [NA]

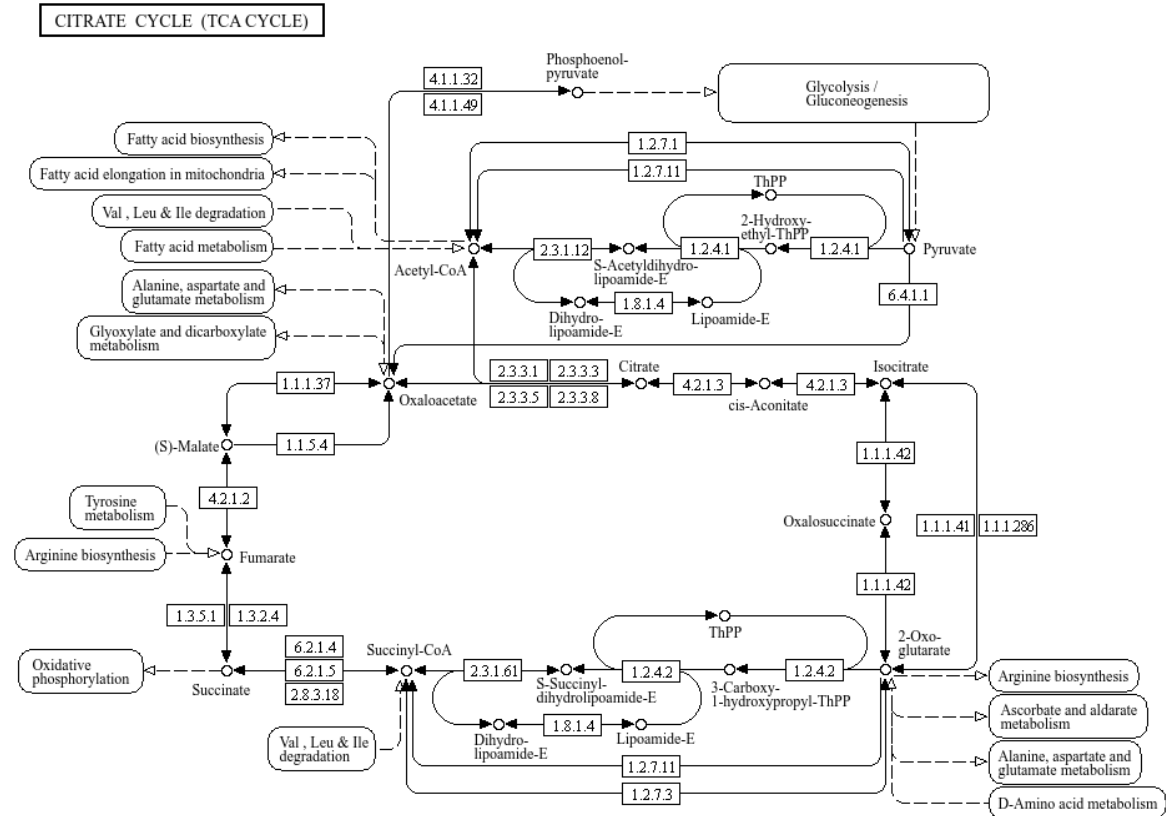
Annotations come with evidence



Pathway databases trace metabolic pathways and their regulation



WIKIPATHWAYS
Pathways for the People



00020 9/27/24
(c) Kanehisa Laboratories



Protein Domain databases annotate functional subdomains within proteins



PH domain

This is a SMART **PH** domain ([full annotation](#)).

Position: 246 to 363

E-value: 9.70401799592535e-12 ([HMMER2](#))

PH

SMART ACC: [SM000233](#)

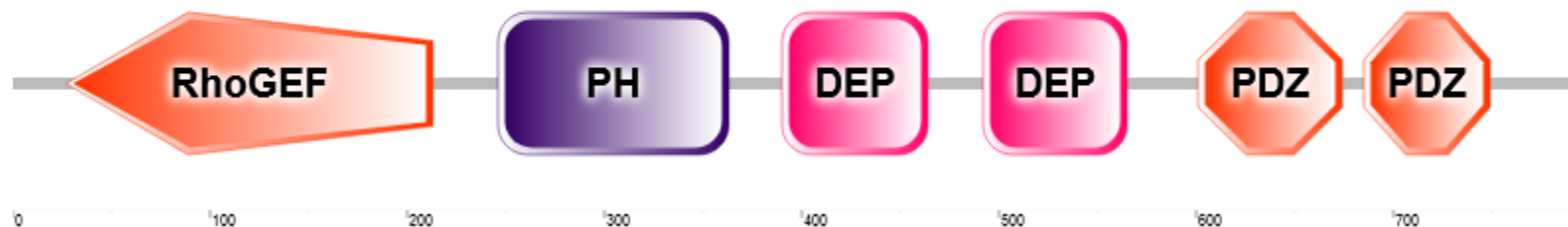
Definition: Pleckstrin homology domain.

Description: Domain commonly found in eukaryotic signalling proteins. The domain family possesses multiple functions including the abilities to bind inositol phosphates, and various proteins. PH domains have been found to possess inserted domains (such as in PLC gamma, syntrophins) and to be inserted within other domains. Mutations in Brutons tyrosine kinase (Btk) within its PH domain cause X-linked agammaglobulinaemia (XLA) in patients. Point mutations cluster into the positively charged end of the molecule around the predicted binding site for phosphatidylinositol lipids.

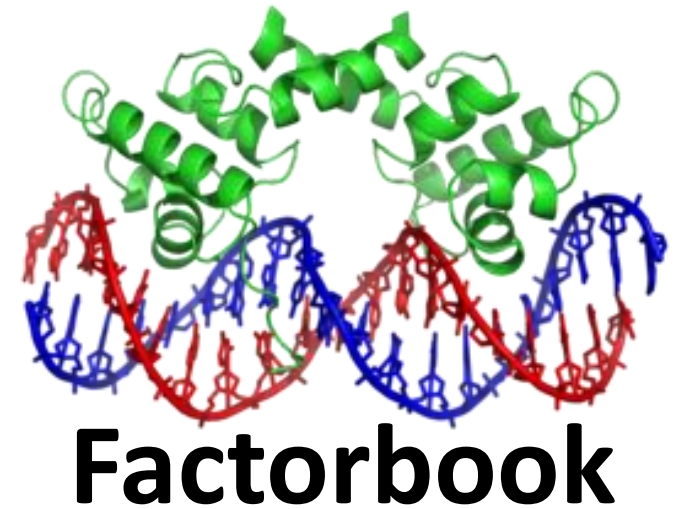
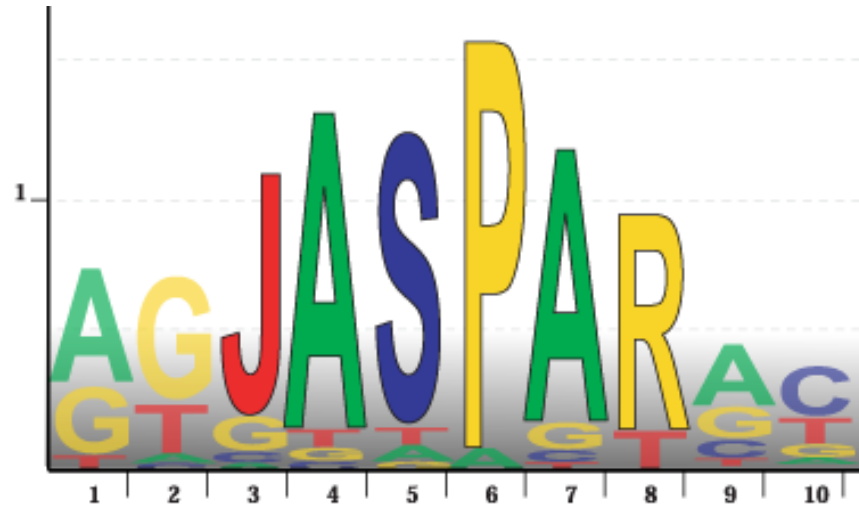
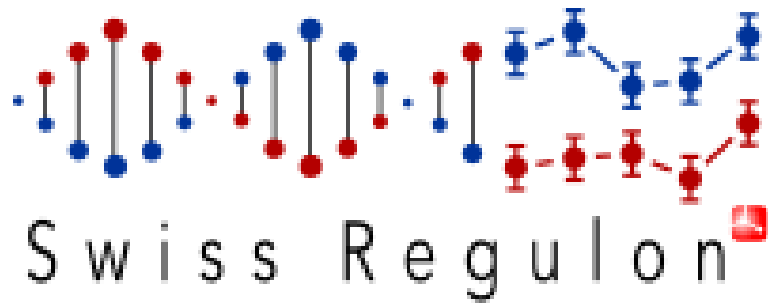
Interpro abstract ([IPR001849](#)): Pleckstrin homology (PH) domains are small modular domains that occur in a large variety of proteins. The domains can bind phosphatidylinositol within biological membranes and proteins such ...([full abstract](#))

Domains within *Homo sapiens* protein **PREX2_HUMAN** (Q70Z35)

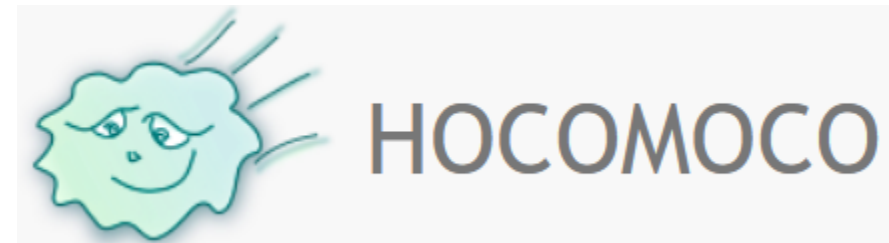
Phosphatidylinositol 3,4,5-trisphosphate-dependent Rac exchanger 2 protein



Transcription Factor databases group genes by the motifs in their promoters



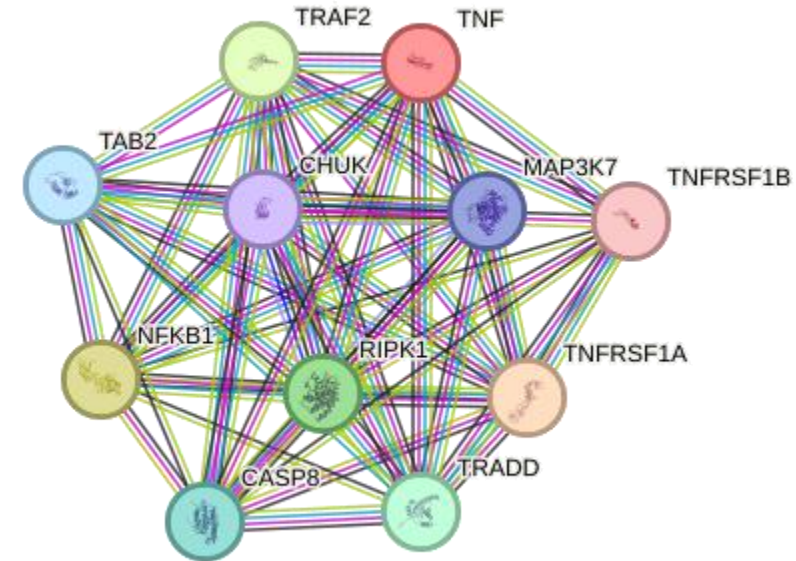
AnimalTFDB v4.0



Interaction databases map out interactions between genes / proteins



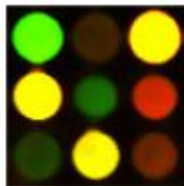
- Physical interaction
- Genetic interaction
- Gene fusions
- Literature mentions
- Genome neighbourhood



Co-expression databases group genes which are expressed together

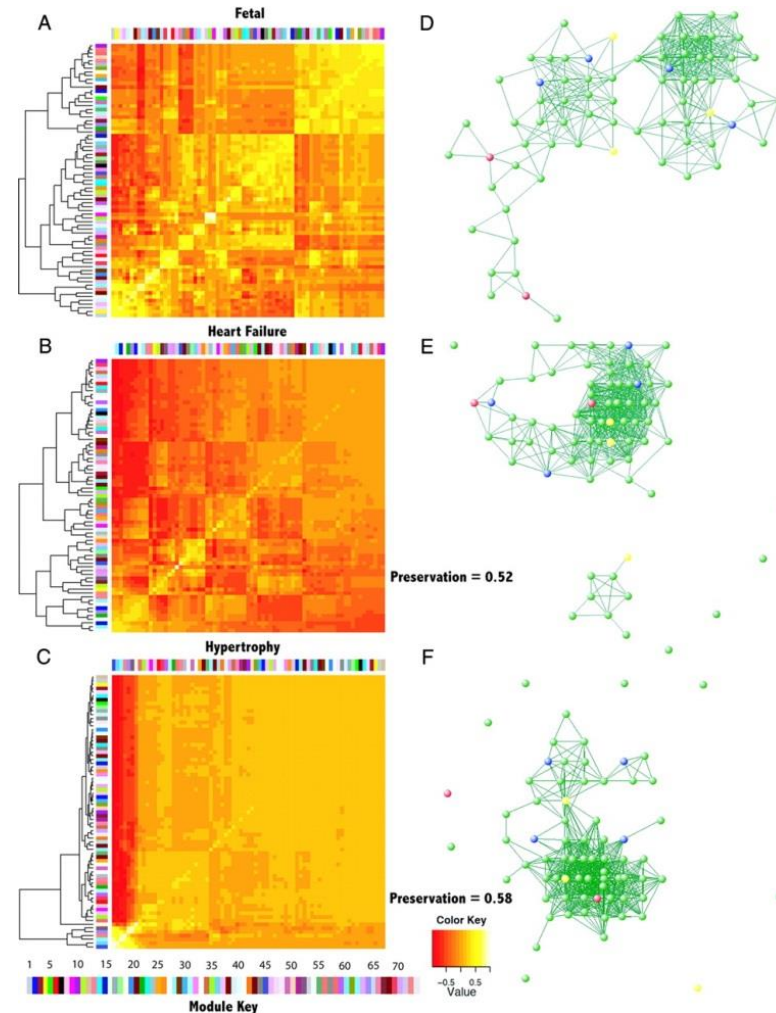


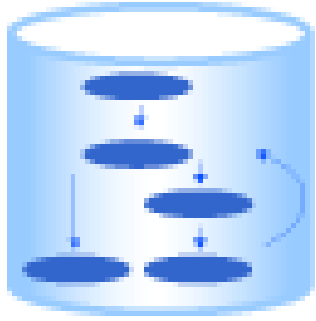
GeneFriends



Coexpedia

Powered by NETBIOLAB.org





MSigDB

Molecular Signatures Database

MH

mouse-ortholog hallmark gene sets are versions of gene sets in the MSigDB Hallmarks collection mapped to their mouse orthologs.

M1

positional gene sets corresponding to mouse chromosome cytogenetic bands.

M2

curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

M3

regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

M5

ontology gene sets consist of genes annotated by the same ontology term.

M8

cell type signature gene sets curated from cluster markers identified in single-cell sequencing studies of mouse tissue.

H

hallmark gene sets are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1

positional gene sets corresponding to human chromosome cytogenetic bands.

C2

curated gene sets from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3

regulatory target gene sets based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4

computational gene sets defined by mining large collections of cancer-oriented expression data.

C5

ontology gene sets consist of genes annotated by the same ontology term.

C6

oncogenic signature gene sets defined directly from microarray gene expression data from cancer gene perturbations.

C7

immunologic signature gene sets represent cell states and perturbations within the immune system.

C8

cell type signature gene sets curated from cluster markers identified in single-cell sequencing studies of human tissue.

<https://www.gsea-msigdb.org/gsea/msigdb>

Testing for enriched gene sets

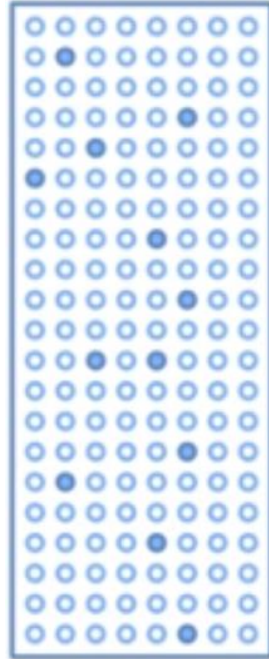
There are two basic ways to test for enrichment

- Categorical
 - Start from a list of hit genes
 - Count overlaps between hit list and functional list
 - Find Functional lists where the degree of overlap is statistically unlikely
- Quantitative
 - Start with all genes
 - Associate a value with each gene
 - Look for functional sets with unusual distributions of values

Categorical Enrichment Analysis

Categorical tests for enrichment

13,101 genes
on chip



3005 genes
related to
disease
 $3005/13,101 =$
23.1%

Gene List



	Gene List	Background
In disease annotated group	260	3005
Not in disease annotated group	487	10096

260



Related to
disease
 $260/747 =$
34.8%

487



Not related to
disease

Fisher's Exact test

	Gene List	Background	Total
In disease annotated group	260 $E = 176.1$	3005 $E = 3088.8$	3265
Not in disease annotated group	487 $E = 570.9$	10096 $E = 10012.1$	10583
Total	747	13101	13848

```
> counts <- (matrix(data = c(260, 487, 3005, 10096), nrow = 2))  
> fisher.test(counts)
```

Fisher's Exact Test for Count Data

data: counts

p-value = 9.769e-13

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

1.52846 2.10120

sample estimates:

odds ratio

1.793564

$(260/487) / (3005/10096)$

Categorical tests are influenced by where you set the cutoff for “interesting” genes

Hit1	Hit17
Hit2	Hit18
Hit3	Hit19
Hit4	Hit20
Hit5	Hit21
Hit6	Hit22
Hit7	Hit23
Hit8	Hit24
Hit9	Hit25
Hit10	Hit26
Hit11	Hit27
Hit12	Hit28
Hit13	Hit29
Hit14	Hit30
Hit15	Hit31
Hit16	Hit32

- Function X
 - 3 hits out of 32 in ‘interesting’ list
 - Not significant ($p=0.07$)

Categorical tests are influenced by where you set the cutoff for “interesting” genes

Hit1	Hit17
Hit2	Hit18
Hit3	Hit19
Hit4	Hit20
Hit5	Hit21
Hit6	Hit22
Hit7	Hit23
Hit8	Hit24
Hit9	Hit25
Hit10	Hit26
Hit11	Hit27
Hit12	Hit28
Hit13	Hit29
Hit14	Hit30
Hit15	Hit31
Hit16	Hit32

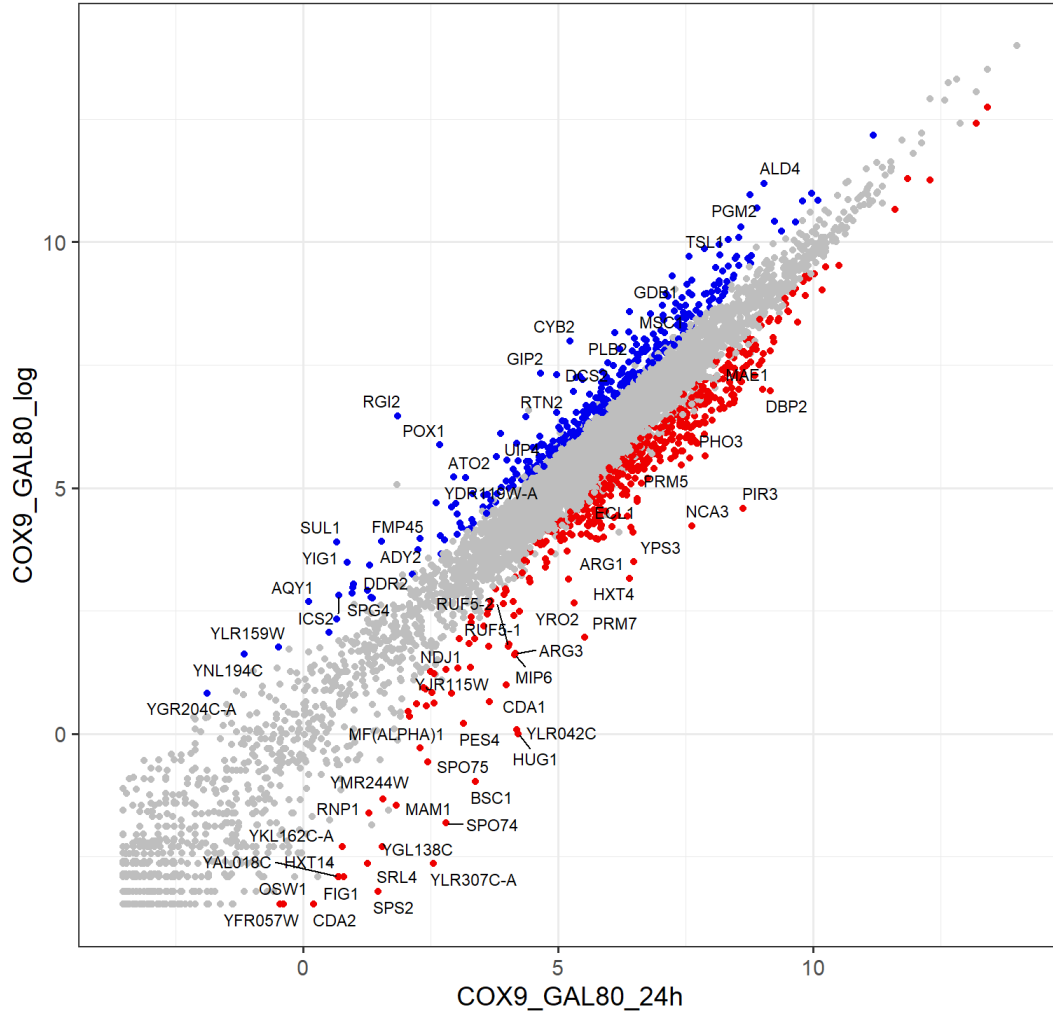
- Function X
 - 3 hits out of 7 in ‘interesting’ list
 - Significant ($p=0.02$)

Ordered, but not quantitative lists allow sequential categorical analysis

Hit1	Hit17
Hit2	Hit18
Hit3	Hit19
Hit4	Hit20
Hit5	Hit21
Hit6	Hit22
Hit7	Hit23
Hit8	Hit24
Hit9	Hit25
Hit10	Hit26
Hit11	Hit27
Hit12	Hit28
Hit13	Hit29
Hit14	Hit30
Hit15	Hit31
Hit16	Hit32

- Function X
 - Length=1 $p=0.60$
 - Length=2 $p=0.80$
 - Length=3 $p=0.30$
 - Length=4 $p=0.35$
 - Length=5 $p=0.40$
 - Length=6 $p=0.45$
 - Length=7 $p=0.05$
 - Length=8 $p=0.08$
 - Length=9 $p=0.10$

Directional Gene Lists

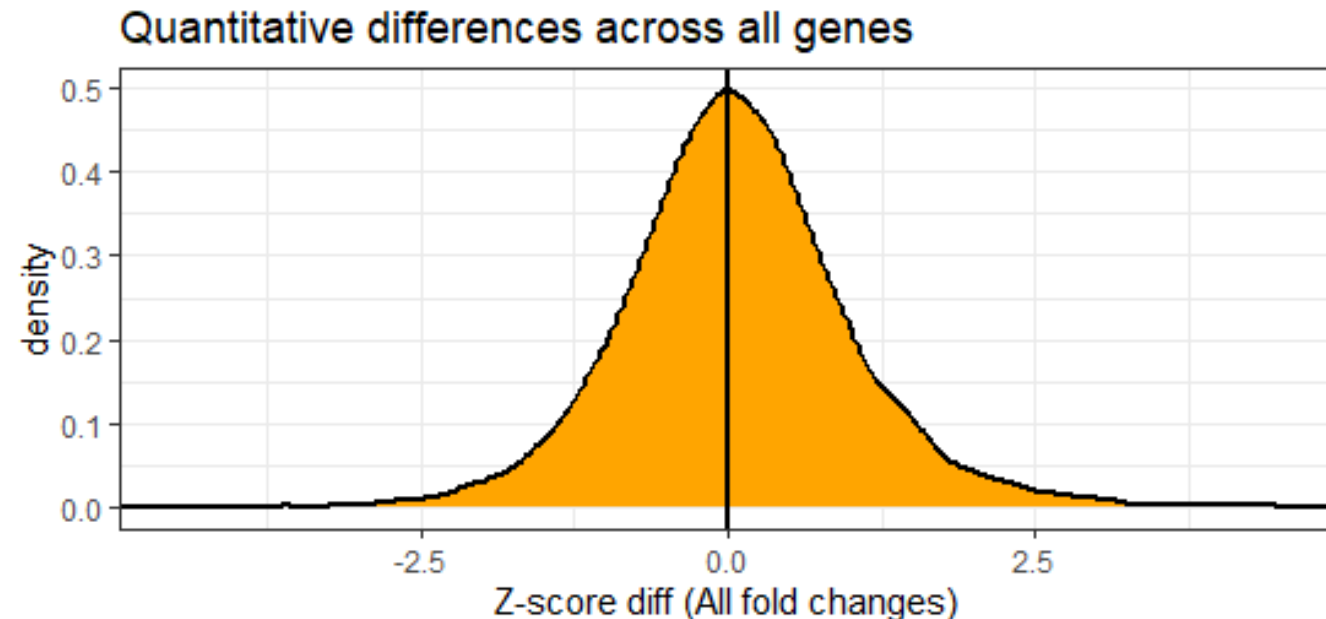


- One search or two?
 - One search
 - Higher power (more genes)
 - Lower enrichment
 - Mixed effects (pathways)
 - Two searches
 - Easier interpretation
 - Less power
 - Higher enrichment

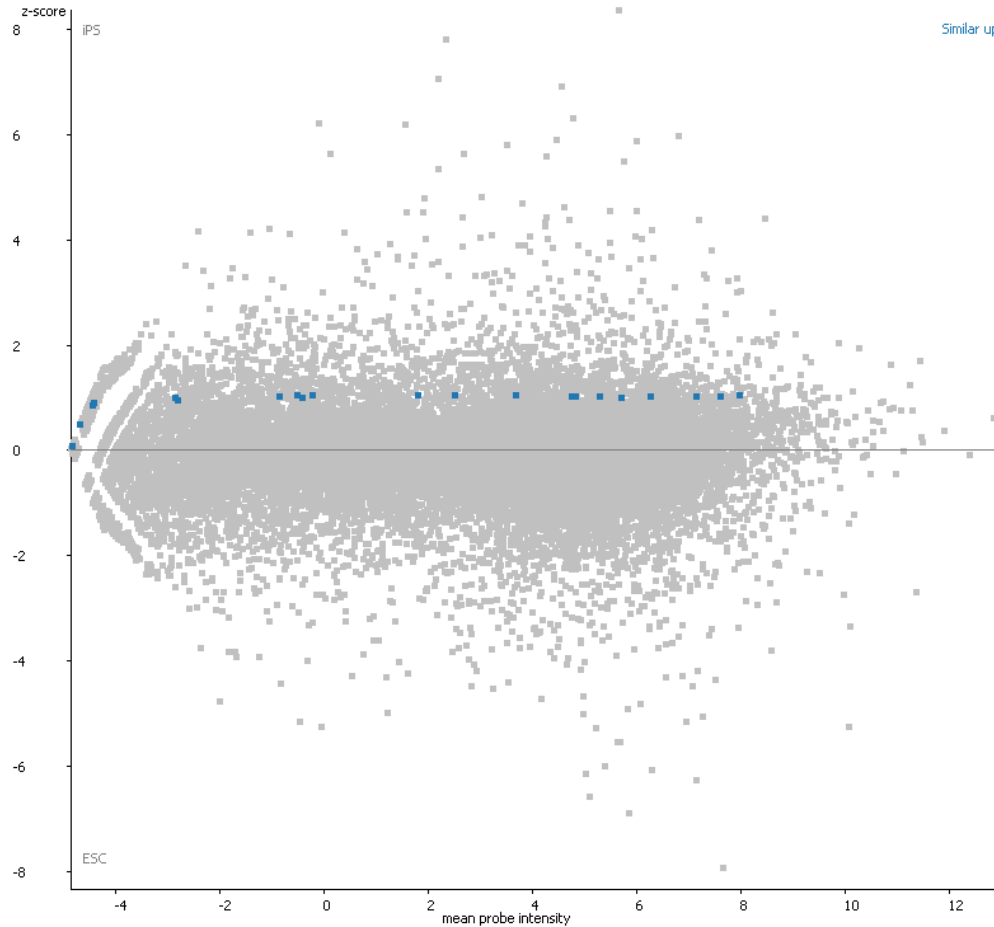
Quantitative Enrichment Analysis

Quantitative comparisons can offer more power

- What quantitative value can we use?
 - Differential p-value (normally $-10 \log(p)$)
 - Fold change
 - Absolute difference



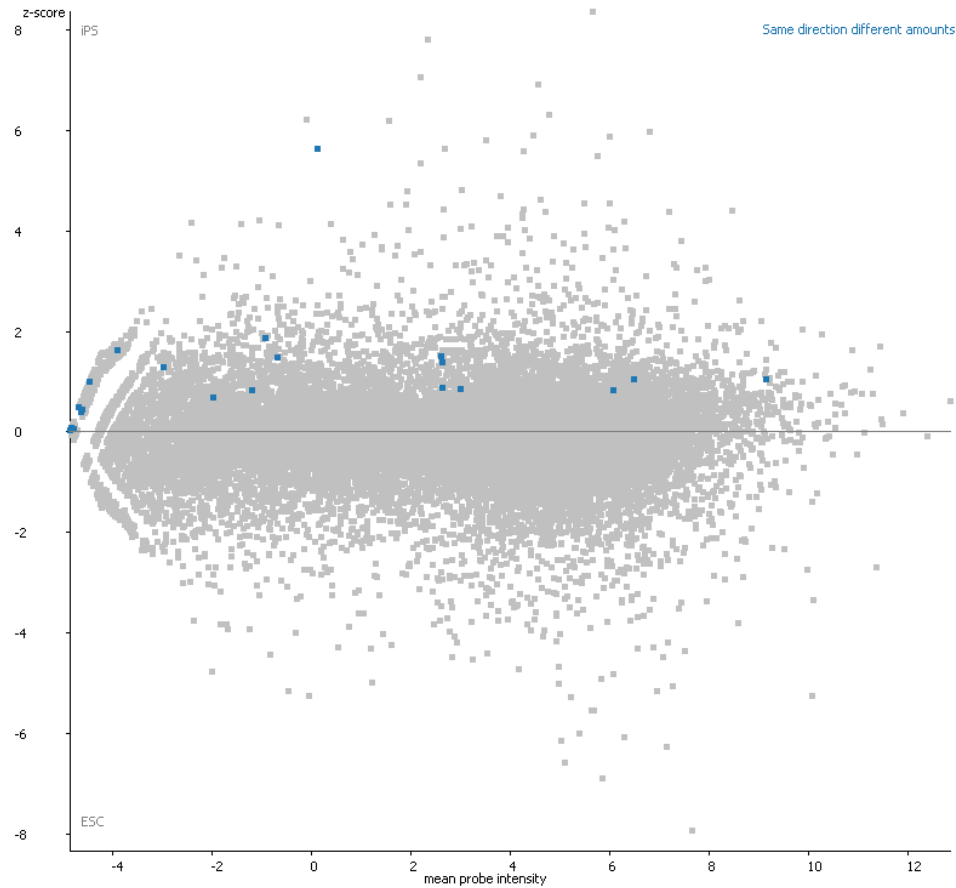
What kind of changes do we expect in an interesting category?



Student's T-test

Genes in that category all change, and by about the same amount?

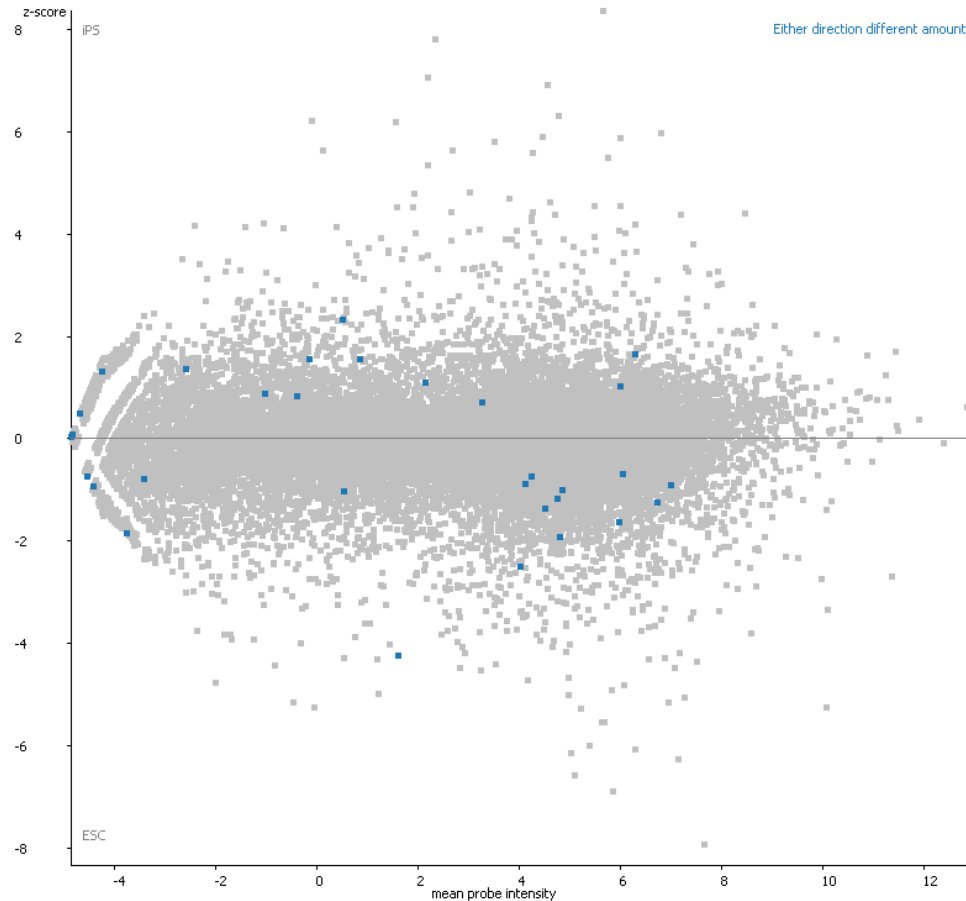
What kind of changes do we expect in an interesting category?



Kolmogorov Smirnov Test

Genes in that category all change in the same direction, but by different amounts?

What kind of changes do we expect in an interesting category?



Absolute KS Test

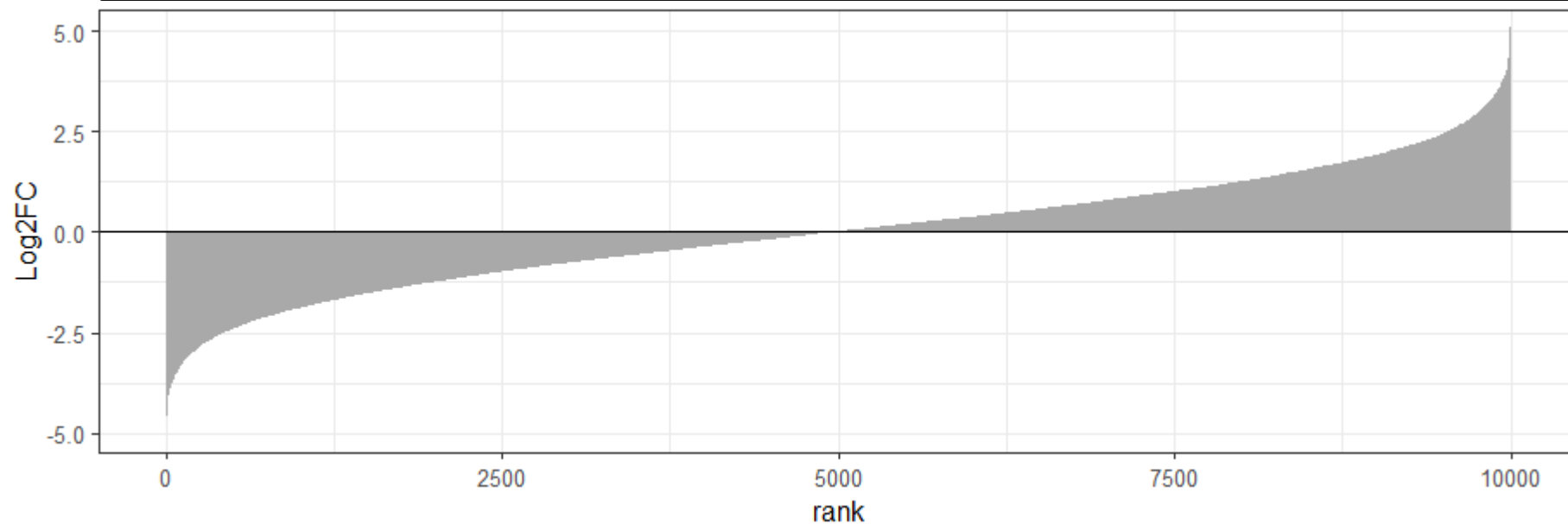
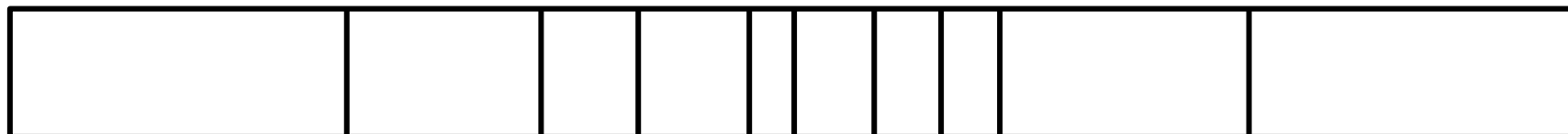
Genes in that category all change in either direction, but by different amounts?

GSEA statistics

Interesting

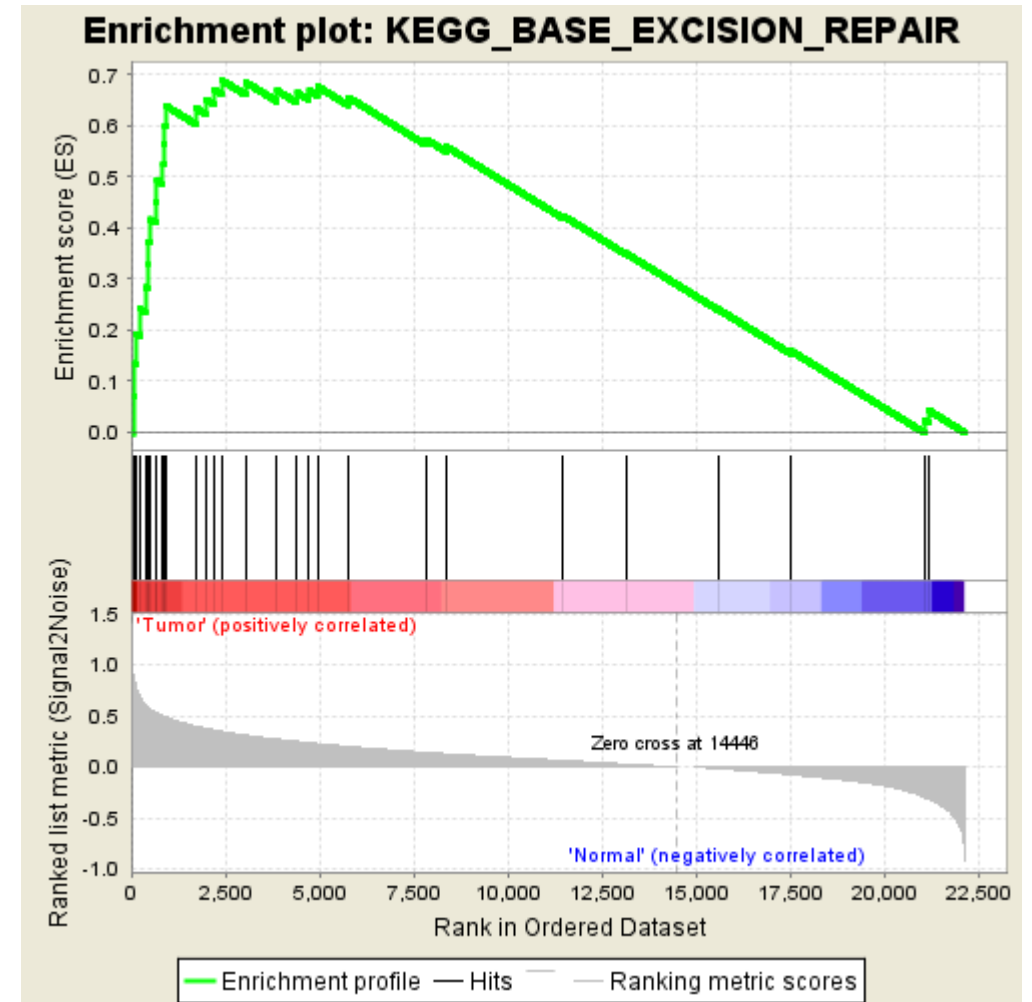


Boring



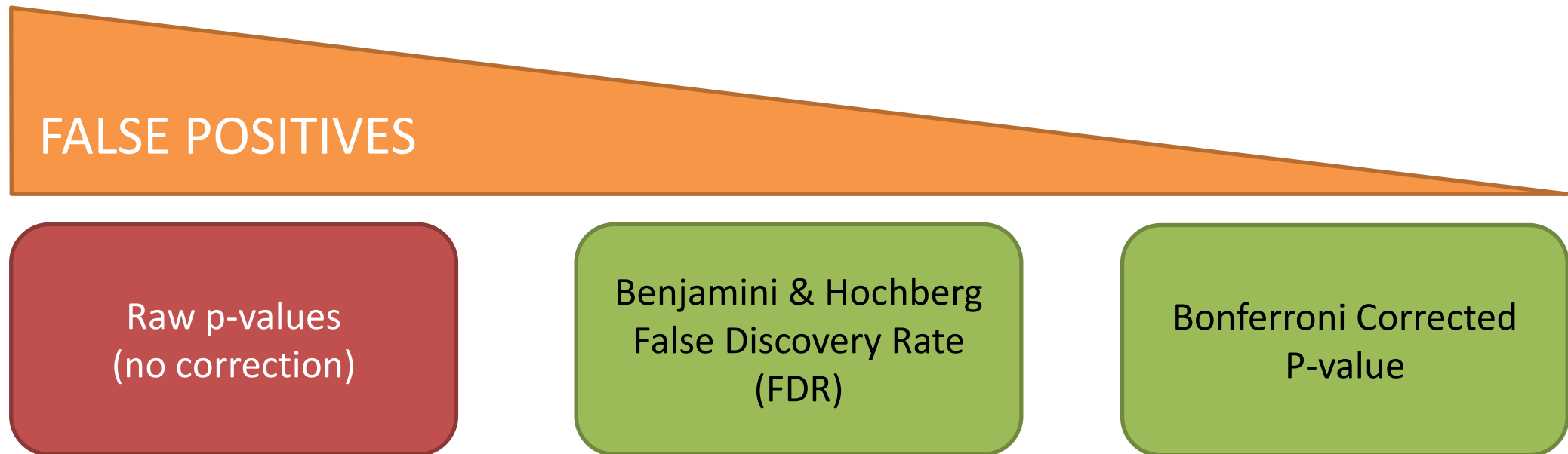
GSEA Statistics

- Keep a running total
- Start at the highest values
- If gene is in the set add value
- Otherwise subtract value
- Enrichment score is max score
- Stats compare ES with randomly shuffled data



Multiple Testing Correction

- Original p-value is for one test (one gene set)
- Thousands of sets tested in each analysis
- Many tools report raw as well as corrected p-values

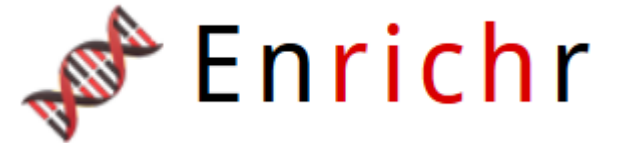


What do we get back from an enrichment test?

- A p-value
 - Remember that this reflects not only difference but also variance and power (number of observations)
- A difference value
 - Enrichment difference (odds ratio)
 - Mean quantitative difference
 - Remember large differences are easier to obtain with small numbers of observations

Tools for functional gene list analysis

- There are many different tools available, both free and commercial
- Popular tools include:



GOliath



GORILLA



Gene Ontology enRIchment anaLysis and visuaLizAtion tool



WebGestalt

g:GOST

Functional profiling

g:Convert

Gene ID conversion

g:Orth

Orthology search

g:SNPense

SNP id to gene name

- Categorical or ordered statistics
- Lots of additional options
- Wide species support
- Interesting presentation
 - Doesn't scale well to lots of hits

Query Upload query Upload bed file

Input is whitespace-separated list of genes ?

Run query random example mixed query example

Options

Organism: ?

Homo sapiens (Human) ▼

☒ Highlight driver terms in GO ?

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼



- Categorical or Quantitative statistics
- Part of Gene Ontology Consortium
 - Annotations are up to date
- Simple enrichment analysis
- Functional lists and categorical break down

1. Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.

Enter IDs: [Supported IDs](#) separate IDs by a space or comma

Upload IDs: [File](#) [format](#) No file selected.

Please [login](#) to be able to select lists from your workspace.

Select List Type: ☒ ID List
☐ Previously exported text search results
☐ Workspace list
☐ PANTHER Generic Mapping
☐ ID's from Reference Proteome Genome

Organism for id list

☐ VCF File Flanking region ☐ Search Enhancer Data



Your data

Options

Analysis

Step 1: Select a file from your computer or paste your own data and click on the corresponding "Continue" button.

- Categorical or quantitative statistics
- Pathway focussed
- Simple submission interface (no custom background)
- Really nice visualisations

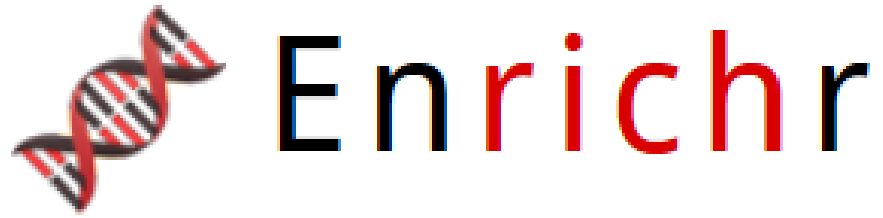
Goliath

- Categorical statistics
- Limited species support
- Allows custom backgrounds
- Uses PathwayCommons gene sets
- Innovative detection and presentation of artefacts

Select species	human
Min Category Size	50
Max Category Size	500
Gene List	Background List (optional)
Paste Gene Names here	Paste Gene Names here
Query name (optional)	



- Categorical Statistics
- Most popular system (mostly historic)
- Has been behind the latest annotation
 - Was updated again, but now behind once more
- Lots of support for different IDs and Species
- Configurable gene sets
- Simple output presentation



- Categorical Statistics
- Biggest selection of gene sets
- Simple interface, but limited options
 - No species information
- Simple interactive visualisation
- Novel scoring scheme to rank hits

Drop a file or paste a set of Entrez gene symbols on each row in the textbox below. You can try a gene set **example**. Also, you can now try adding a **background** (**clear**).

Drop a file or paste a set of valid Entrez gene symbols (e.g. STAT3) on each row in the text-box

A1BG
A2M
NAT1
NAT2
SERPINA3
AADAC
AAMP
AANAT
AARS1
ABAT

0 gene(s) entered



GO RILLA



Gene Ontology enRIchment anaLysis and visualiZAtion tool

- Categorical or ranked analysis
- Mostly GO gene list support
- Interesting visualisation options

Step 1: Choose organism

Homo sapiens ▼

Step 2: Choose running mode

☐ Single ranked list of genes ☒ Two unranked lists of genes (target and background lists)

Step 3: Paste a ranked list of gene/protein names

Names should be separated by an <ENTER>. The preferred format is gene symbol. Other supported formats are: gene and protein RefSeq, Uniprot, Unigene and Ensembl.

Target set:

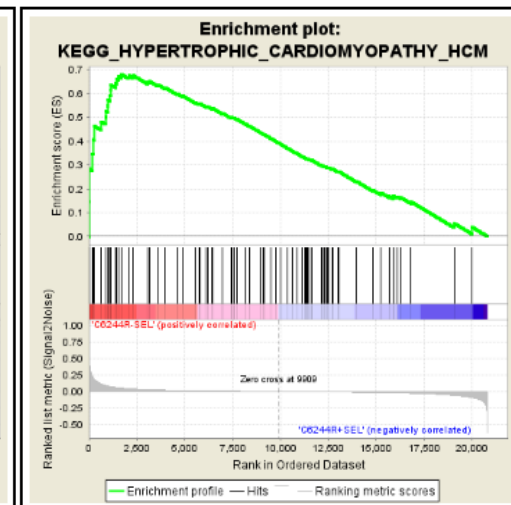
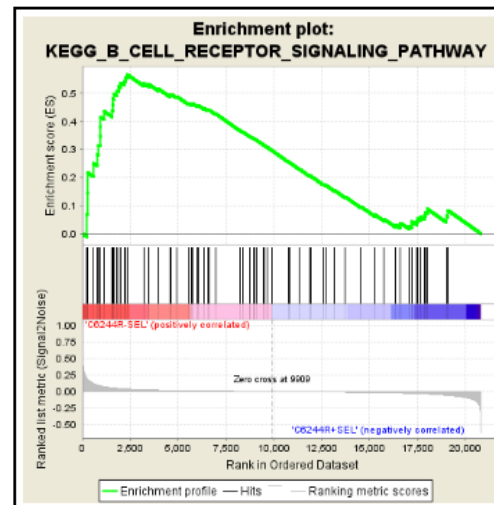
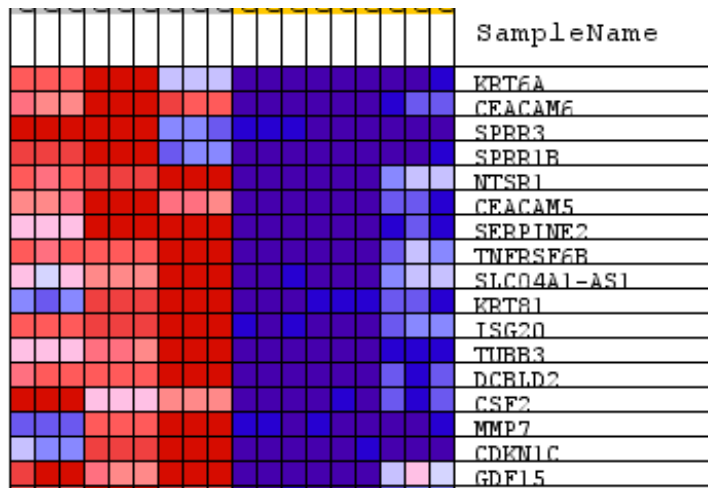
Or upload a file: No file selected.

Background set:

Or upload a file: No file selected.

GSEA

- Quantitative enrichment
- Designed for expression datasets
- Local application
- Imports tab delimited expression data





SeqMonk Mapped Sequence Data Analyser

Version: 1.48.2.devel

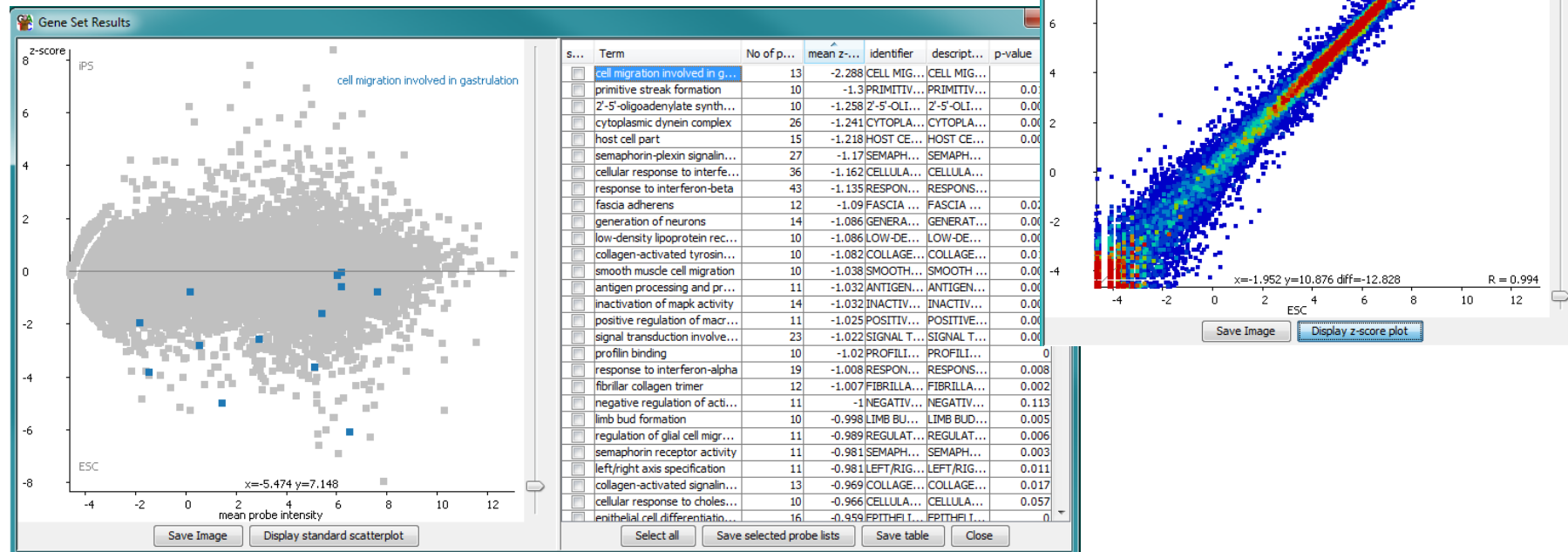
www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews, Laura Biggins Babraham Bioinformatics, 2006-21

HTSJDK BAM/SAM reader ©The Broad Institute, 2009-19



- Quantitative enrichment of sequencing datasets
- Local Java application



Gene List Practical

<https://tinyurl.com/exercisetostartat>

Exploring and Presenting Results

Simon Andrews, Laura Biggins

simon.andrews@babraham.ac.uk

laura.biggins@babraham.ac.uk



Functional enrichment results

- Gene set information
 - Gene set name
 - Gene set source
 - Gene set description
- Count information
 - Hit genes in category
 - Hit genes outside category
 - Background genes in category
 - Background genes outside category
- Statistical information
 - Raw p-value
 - Corrected p-value
 - Enrichment value

Functional enrichment results

- Gene set information
 - Gene set name
 - Gene set source
 - Gene set description
- Count information
 - Hit genes in category
 - Hit genes outside category
 - Background genes in category
 - Background genes outside category
- Statistical information
 - Raw p-value
 - Corrected p-value
 - Enrichment value

Tables are often enough

Category	<i>P</i> value	Genes in GO category over-expressed	% of differentially expressed genes in GO category	Genes in GO category on array	% genes on array in GO category
Over-expressed in AJC:					
Biological process					
GO:9058: biosynthesis	0.009	52	24.41	1264	17.82
GO:7610: behavior	0.019	10	4.695	156	2.2
Over-expressed in AJC:					
Molecular function					
GO:5198: structural molecule activity	< 0.001	43	17.92	750	9.043
Over-expressed in SL:					
Biological process					
GO:8152: metabolism	0.019	192	71.91	4674	65.91
Over-expressed in SL:					
Molecular function					
GO:16209: antioxidant activity	0.013	6	1.917	52	0.627
GO:8135: translation factor activity, nucleic acid binding	0.010	13	4.153	166	2.001
GO:45182: translation regulator activity	0.014	13	4.153	173	2.086
GO:5489: electron transporter activity	0.005	17	5.431	225	2.713
GO:8233: peptidase activity	0.043	28	8.946	529	6.378
GO:3824: catalytic activity	0.001	166	53.04	3645	43.95

These are the significant GO Slim categories representing both biological process and molecular function ontologies for population specific significantly over-expressed ($P \leq 0.05$; no multiple test correction) features. For each significant GO category, we include the *P* value number of over-expressed genes in that GO, percentage of representation in the over-expressed list, number of features of that GO in the microarray, and percentage of representation on the entire microarray.

	Gene Ontology Term	% ¹	Univariate p-value ²	FDR-adjusted p-value ³
B/P Cluster	Immunoglobulin	34.5	4.6E-25	1.8E-23
	Immunoglobulin V-set	37.9	2.6E-18	2.5E-17
	Antigen binding	27.6	8.7E-16	1.2E-14
	Immunoglobulin-like fold	44.8	9.7E-16	4.8E-15
	Immune response	41.4	2.0E-13	2.6E-11
T/NK Cluster	Positive regulation of immune system process	24.4	1.7E-08	2.5E-05
	Natural killer cell mediated cytotoxicity	19.5	9.7E-07	5.9E-05
	Positive regulation of lymphocyte activation	17.1	3.3E-07	6.9E-05
	T-cell	12.2	1.3E-06	7.2E-05
	Positive regulation of lymphocyte differentiation	12.2	3.7E-06	3.3E-04
M/D Cluster	MHC class II, alpha/beta chain, N-terminal	39.1	7.0E-22	3.4E-20
	Class II histocompatibility antigen	39.1	1.3E-19	1.9E-18
	MHC class II protein complex	39.1	4.6E-20	3.3E-18
	Immunoglobulin C1-set	43.5	7.0E-18	1.7E-16
	Antigen processing and presentation	47.8	1.3E-18	3.0E-16

¹ percentage of cluster genes (relative to all genes on array) annotated for a given ontology term;

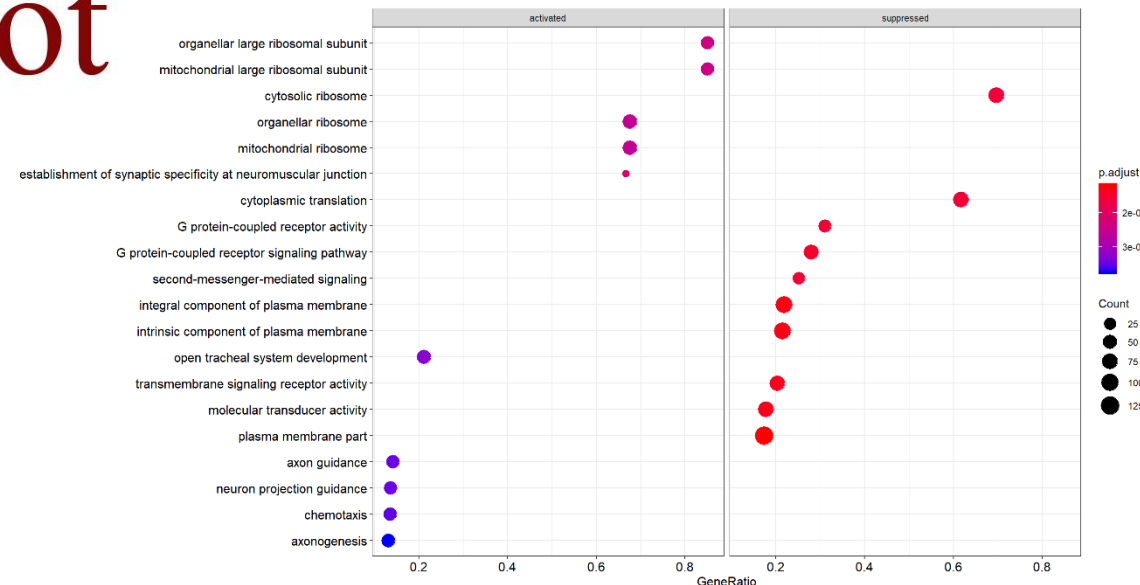
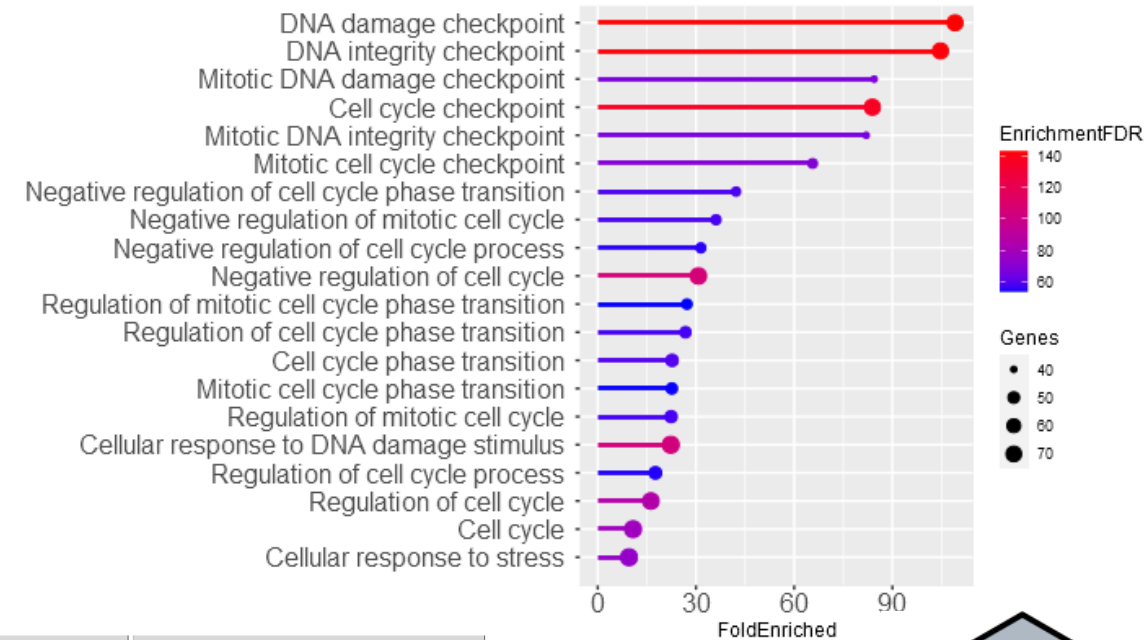
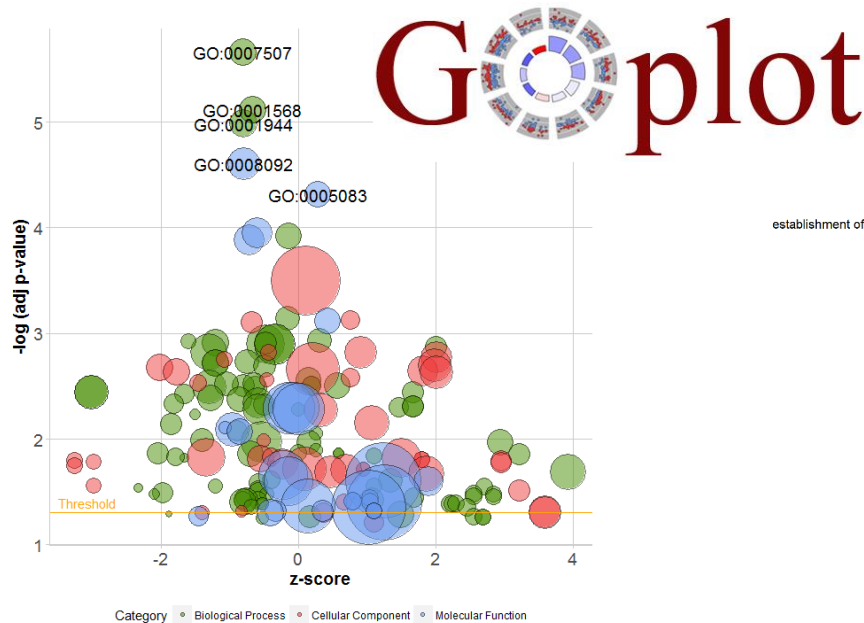
² modified Fisher's Exact Test; ³ Benjamini and Hochberg false discovery rate-adjusted *p* value

Graphical Representations

- Need to add something over a table
 - Relationships between multiple result values
 - Representation of redundancy between categories
 - Relationship to original data
 - Context of surrounding pathway

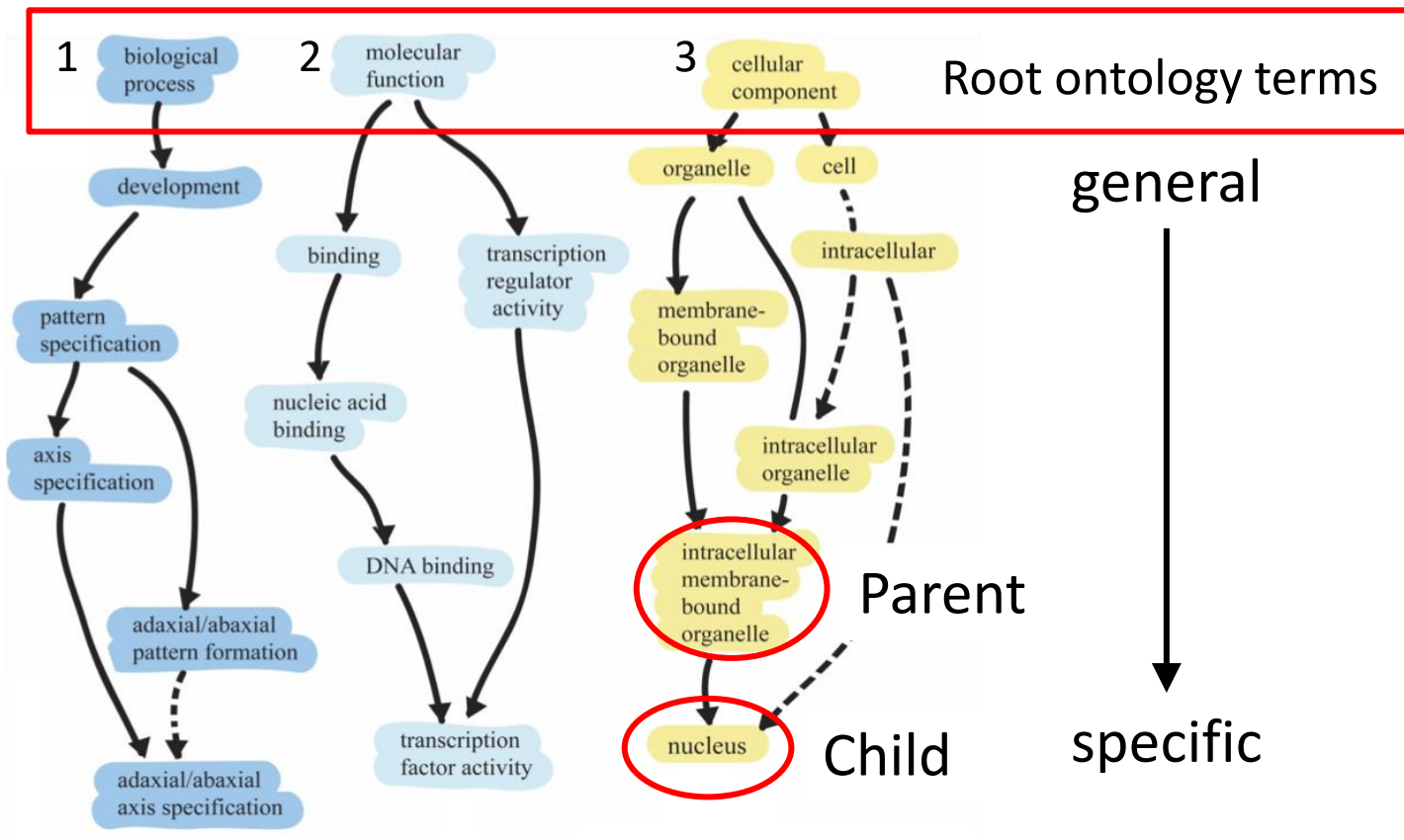
Plotting relationships between values

- P-value(corrected)
- Enrichment
- Size of gene set



Redundancy in gene lists

Gene ontology is hierarchical - a gene is placed in the most specific category and will also appear in all the parent categories

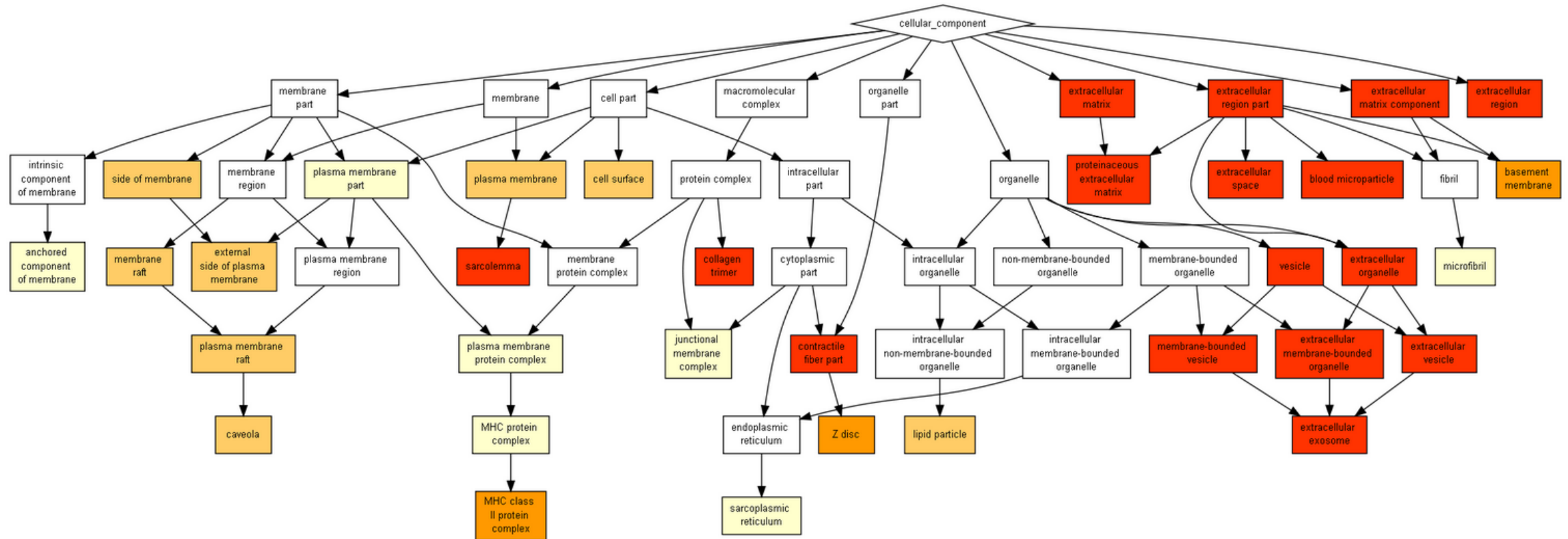


Redundancy: DAVID clustering

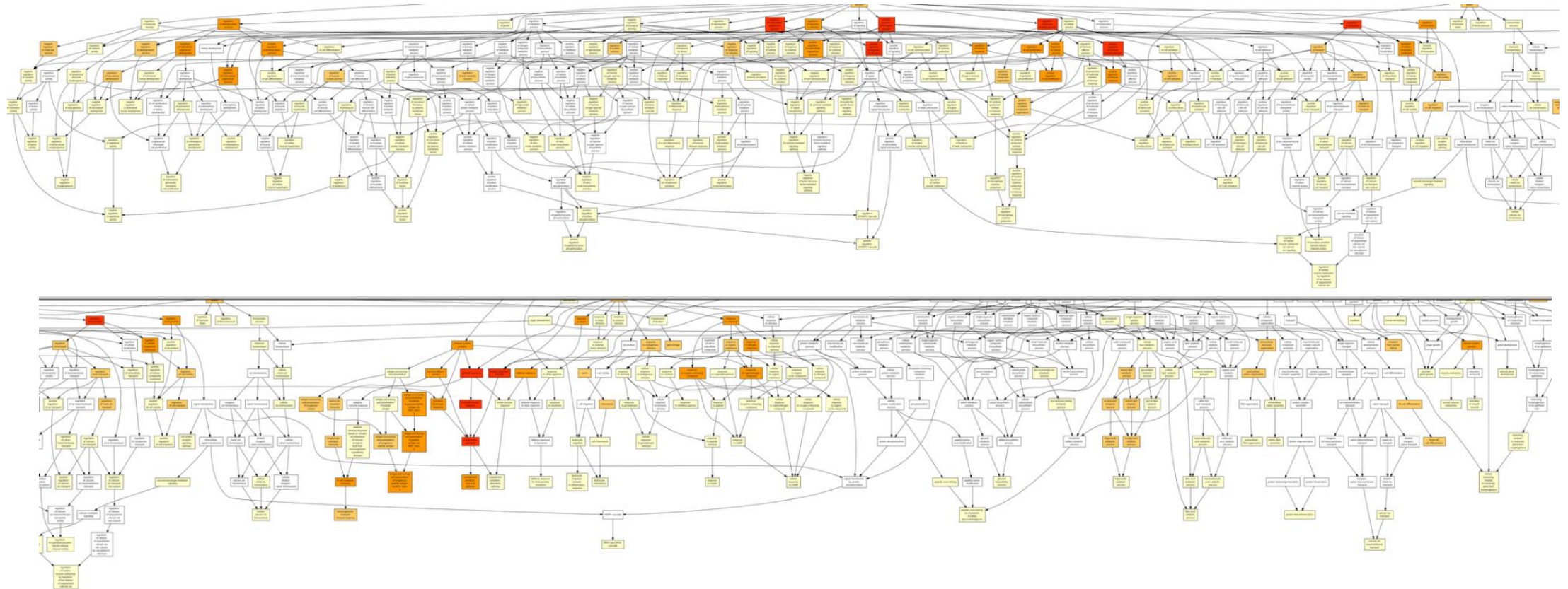
40 Cluster(s) [Download File](#)

Annotation Cluster 1		Enrichment Score: 16.36			Count	P_Value	Benjamini
<input type="checkbox"/>	SP_PIR_KEYWORDS	dna-binding	RT		53	3.5E-24	4.5E-22
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of transcription	RT		60	2.0E-20	1.8E-17
<input type="checkbox"/>	GOTERM_MF_FAT	DNA binding	RT		54	6.0E-20	7.9E-18
<input type="checkbox"/>	GOTERM_MF_FAT	transcription regulator activity	RT		45	2.0E-19	1.3E-17
<input type="checkbox"/>	SP_PIR_KEYWORDS	transcription regulation	RT		49	5.9E-19	3.8E-17
<input type="checkbox"/>	GOTERM_MF_FAT	sequence-specific DNA binding	RT		30	1.5E-16	4.9E-15
<input type="checkbox"/>	SP_PIR_KEYWORDS	Transcription	RT		48	8.1E-16	3.3E-14
<input type="checkbox"/>	GOTERM_BP_FAT	transcription	RT		48	1.9E-15	8.1E-13
<input type="checkbox"/>	GOTERM_MF_FAT	transcription factor activity	RT		33	2.8E-15	9.1E-14
<input type="checkbox"/>	SP_PIR_KEYWORDS	nucleus	RT		69	1.1E-14	3.6E-13
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of RNA metabolic process	RT		40	2.1E-12	6.1E-10
<input type="checkbox"/>	GOTERM_BP_FAT	regulation of transcription, DNA-dependent	RT		39	6.4E-12	1.4E-9
Annotation Cluster 2		Enrichment Score: 10.03			Count	P_Value	Benjamini
<input type="checkbox"/>	GOTERM_MF_FAT	sequence-specific DNA binding	RT		30	1.5E-16	4.9E-15
<input type="checkbox"/>	GOTERM_MF_FAT	transcription factor activity	RT		33	2.8E-15	9.1E-14
<input type="checkbox"/>	INTERPRO	Homeodomain-related	RT		16	2.3E-10	4.3E-8
<input type="checkbox"/>	INTERPRO	Homeobox	RT		15	1.8E-9	1.7E-7
<input type="checkbox"/>	INTERPRO	Homeobox, conserved site	RT		14	3.4E-9	2.1E-7
<input type="checkbox"/>	SP_PIR_KEYWORDS	Homeobox	RT		15	8.5E-9	1.8E-7
<input type="checkbox"/>	UP_SEQ_FEATURE	DNA-binding region:Homeobox	RT		13	2.6E-8	3.7E-6
<input type="checkbox"/>	SMART	HOX	RT		15	4.7E-8	2.1E-6
Annotation Cluster 3		Enrichment Score: 5.86			Count	P_Value	Benjamini
<input type="checkbox"/>	INTERPRO	Transcription factor, fork head, conserved site	RT		7	3.6E-7	1.7E-5
<input type="checkbox"/>	INTERPRO	Transcription factor, fork head	RT		7	3.6E-7	1.7E-5
<input type="checkbox"/>	UP_SEQ_FEATURE	DNA-binding region:Fork-head	RT		7	9.1E-7	6.5E-5
<input type="checkbox"/>	SMART	FH	RT		7	1.8E-6	4.0E-5
<input type="checkbox"/>	INTERPRO	Winged helix repressor DNA-binding	RT		9	2.5E-5	6.6E-4

Redundancy: Gorilla GO images



Redundancy: Gorilla GO images



Enriched Terms

glycogen biosynt...
mitochondrion or...
maturation of 5,8...
mitochondrial cal...
generation of pre...
retinal metabolic...
retinoic acid met...
protein polyubiqu...
pentose-phospho...
SCF-dependent...
heme biosynthesi...
carbohydrate ph...
cellular nitrogen...
molybdenum co...
calcium-mediate...

Input Genes

PSMB1
FBXO9
PSMC6
FBXL3
YME1L1
KDR
GBE1
GYS2
PGM2
DHTKD1
COX15
FECH
RBKS
CAT
GNMT
ALDH8A1
ALDH1A3
GPHN
C1D
EXOSC4

GO

☐

☒

☐

☒

☒

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☒

☐

☐

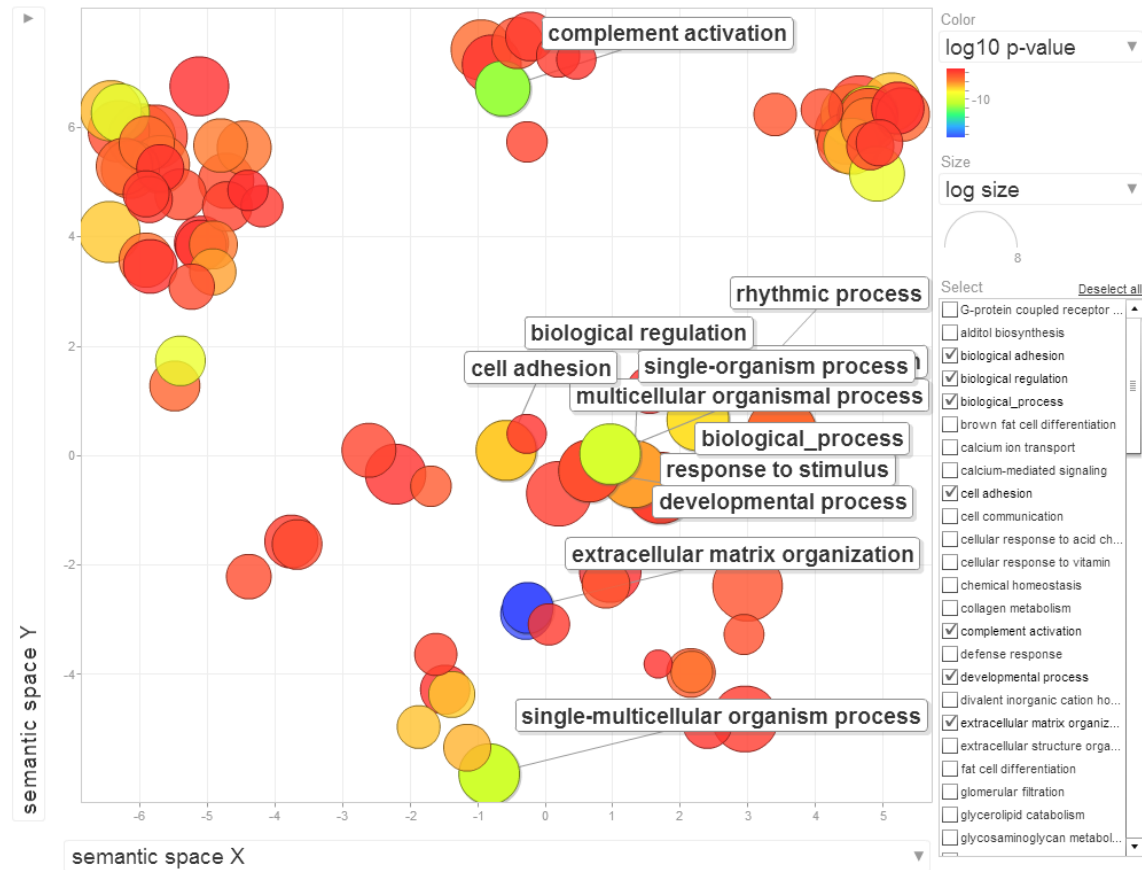


Enrichr

g:Profiler

[illegible]

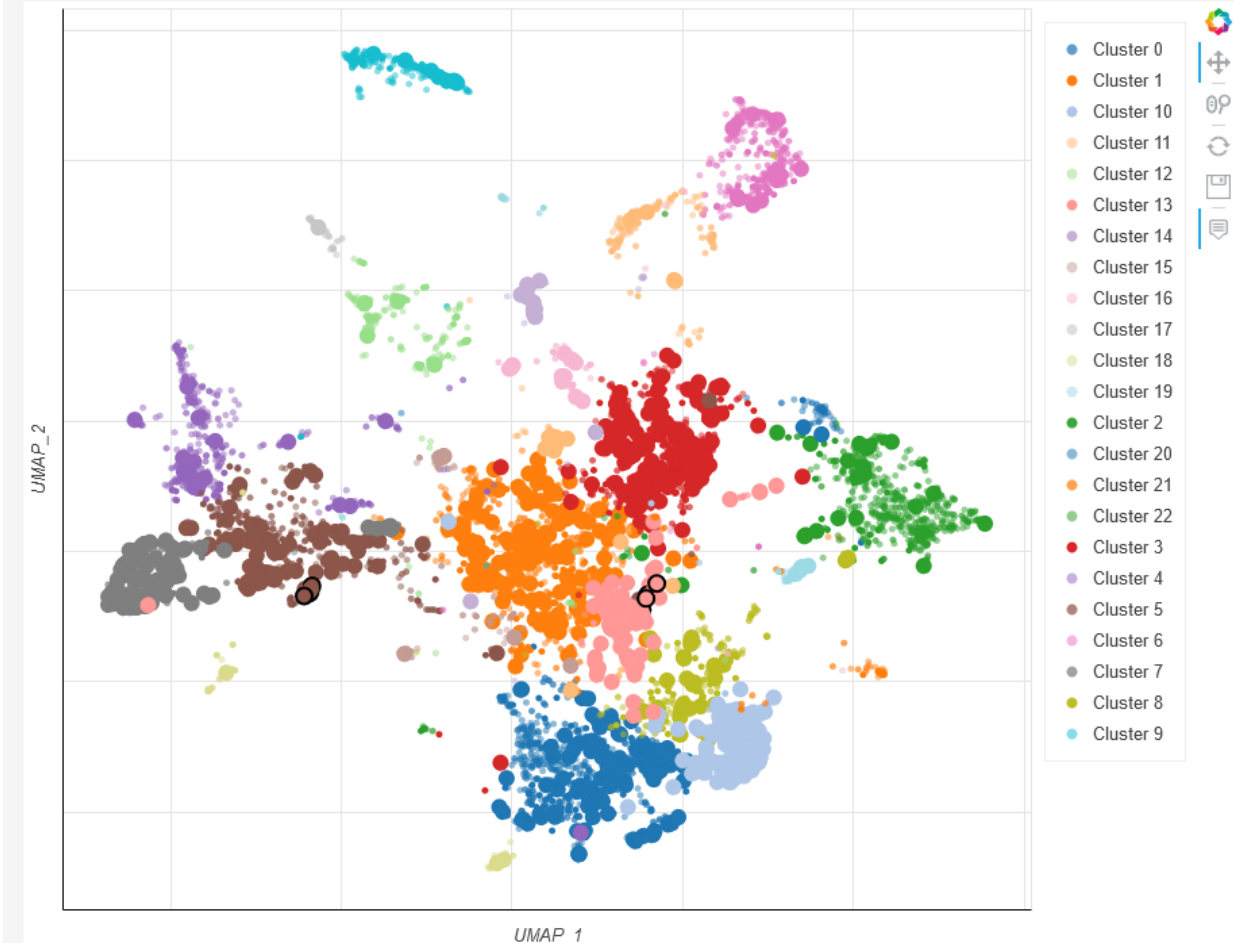
2D Redundancy



Revigo (from Gorilla)
<http://revigo.irb.hr/>

Enrichment Analysis Visualisation (from Enrichr)

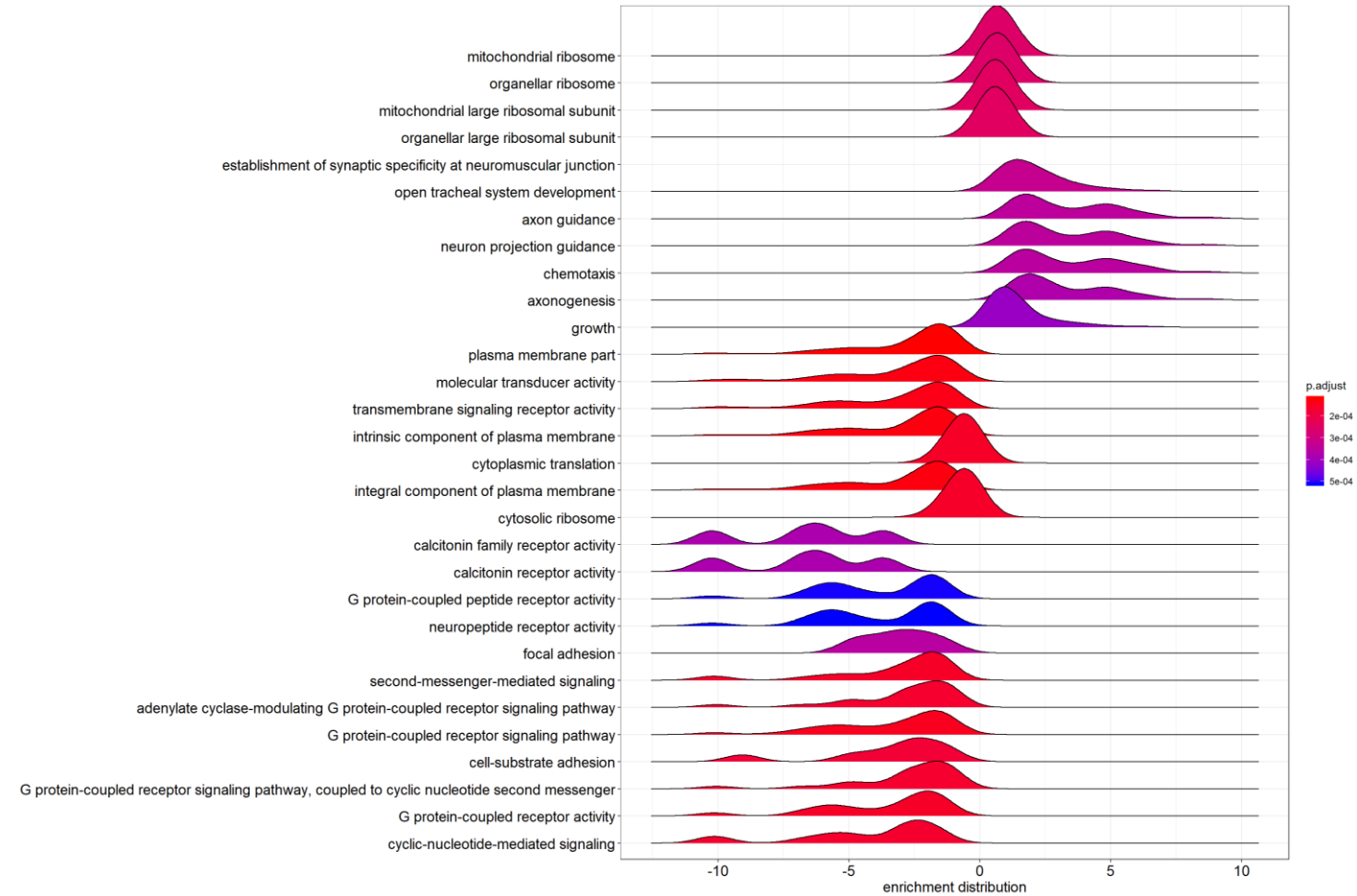
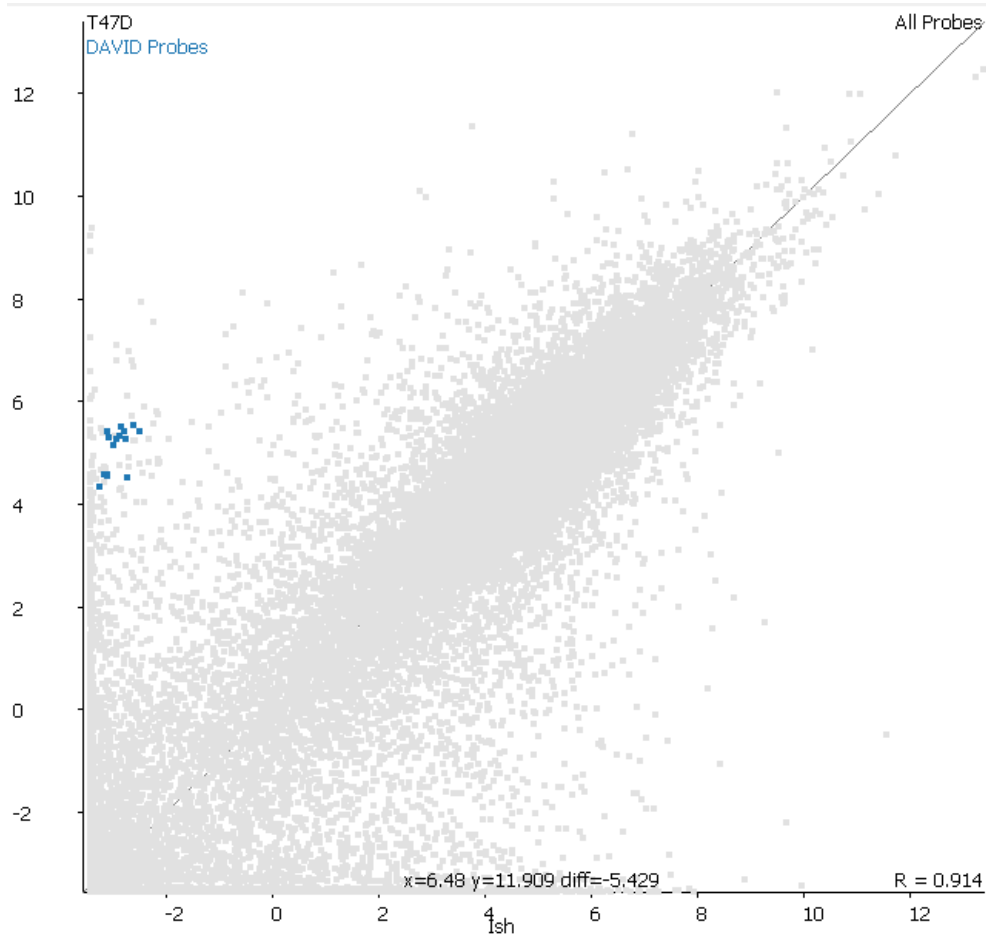
https://appymers.maayanlab.cloud/Enrichment_Analysis_Visualizer/
Scatter plot visualization for GO_Biological_Process_2023.



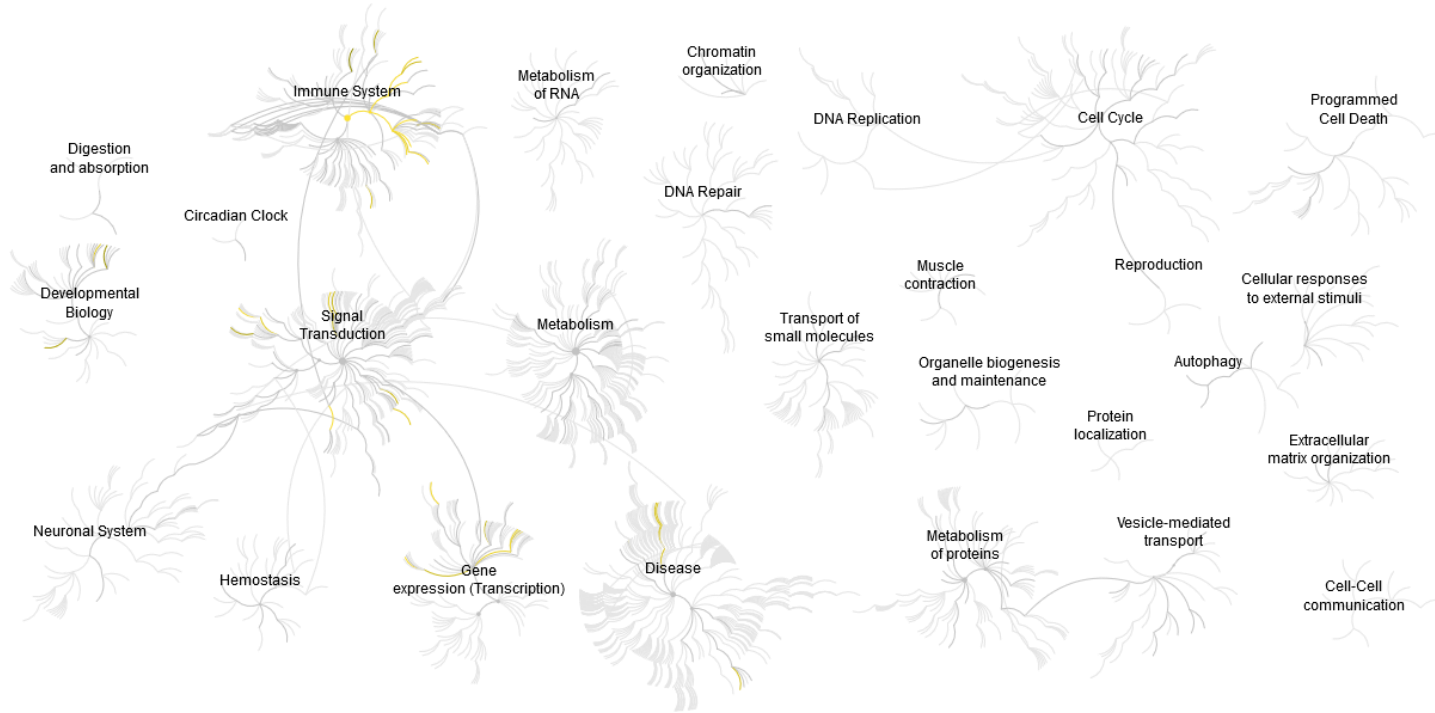
Relationship to original data

- Quantitative values for genes in category
 - Direction and magnitude of change
- Look at genes in category which aren't hits
 - Relative numbers
 - Supportive changes?

Relationship to original data

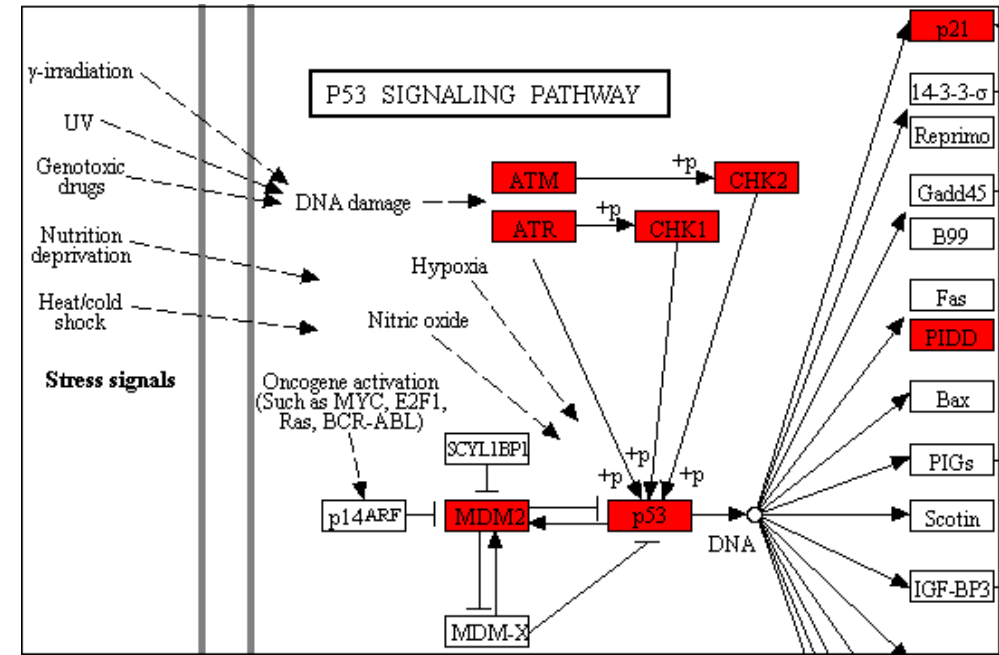


Pathways

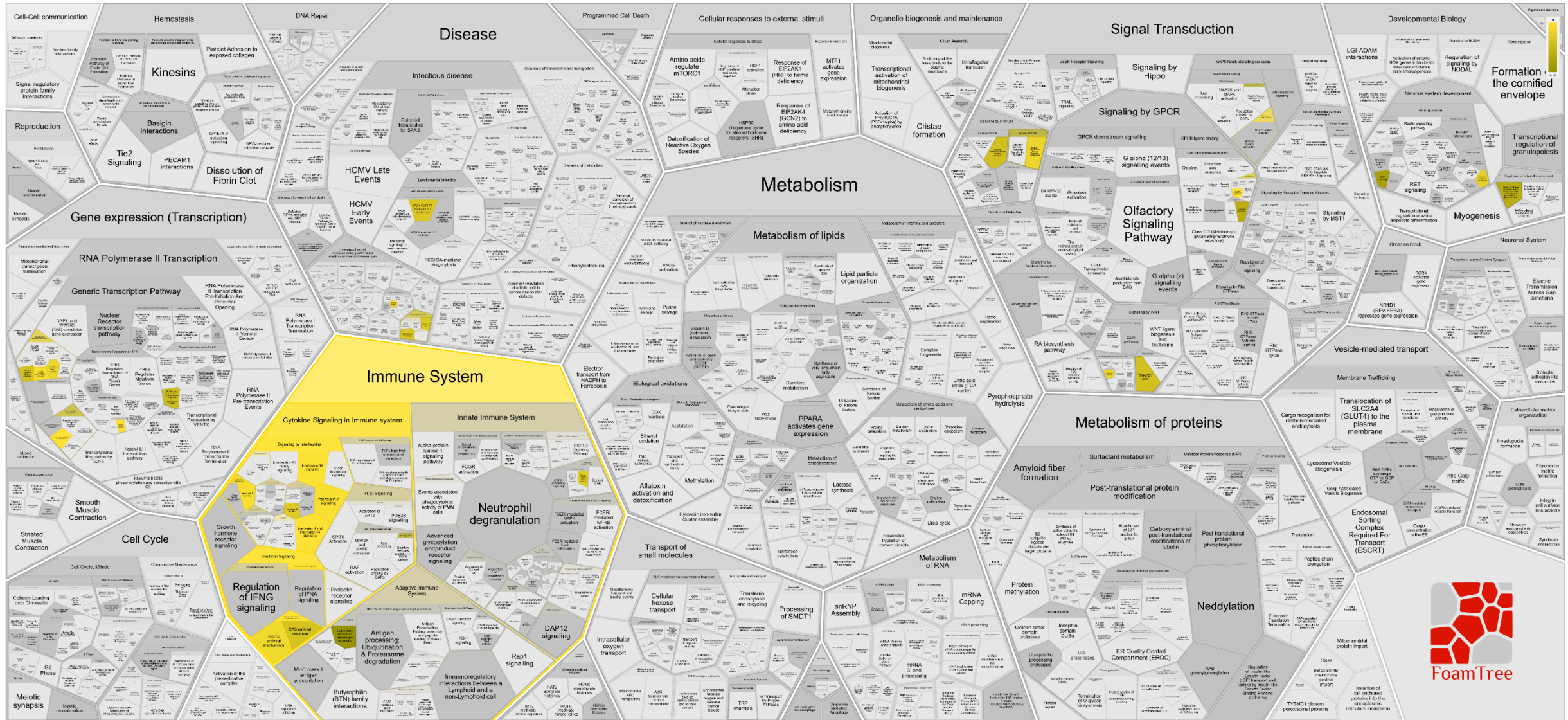


Reactome

ShinyGO



Pathways: Reactome



Summary

- Tables are often sufficient
 - Must include name, enrichment, corrected p-value
 - Other values are useful, but don't put in everything
- Figures can add extra information
 - Plotting multiple metrics
 - Illustrating redundancy
 - Relating to original data
 - Mapping to pathways

Artefacts and Biases in Gene Set Analysis

Simon Andrews, Laura Biggins, Christel Krueger

simon.andrews@babraham.ac.uk

laura.biggins@babraham.ac.uk

christel.krueger@altoslabs.com

What does gene set enrichment test?

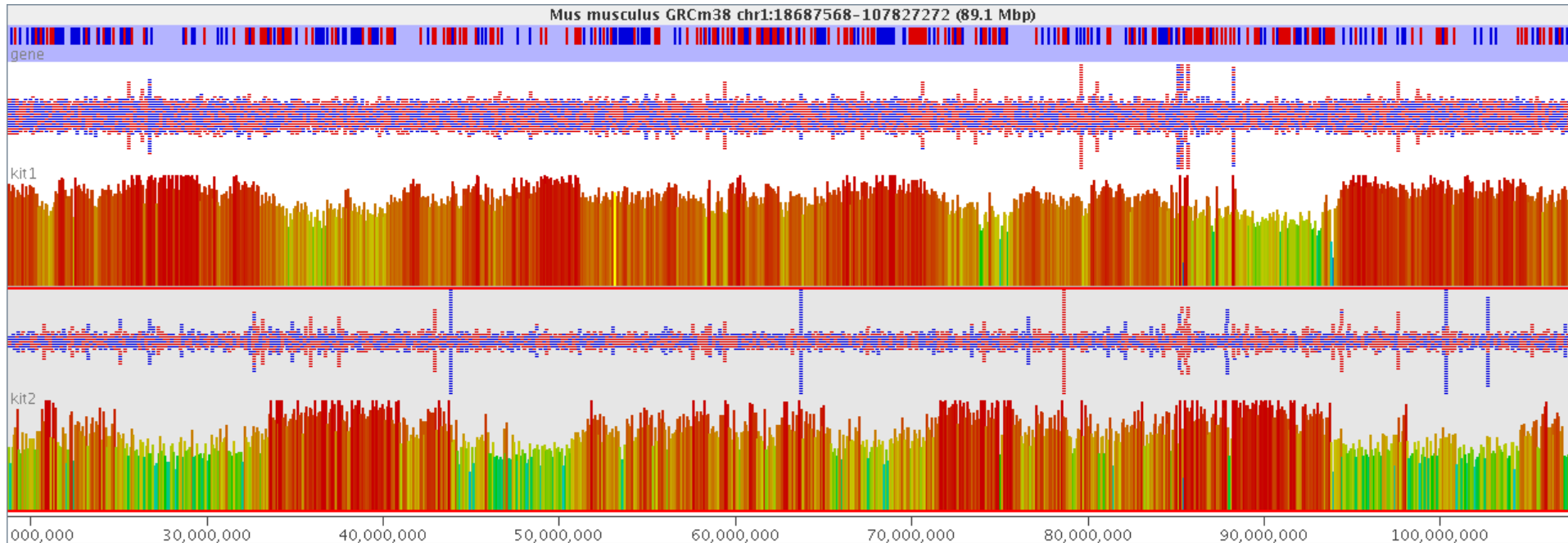
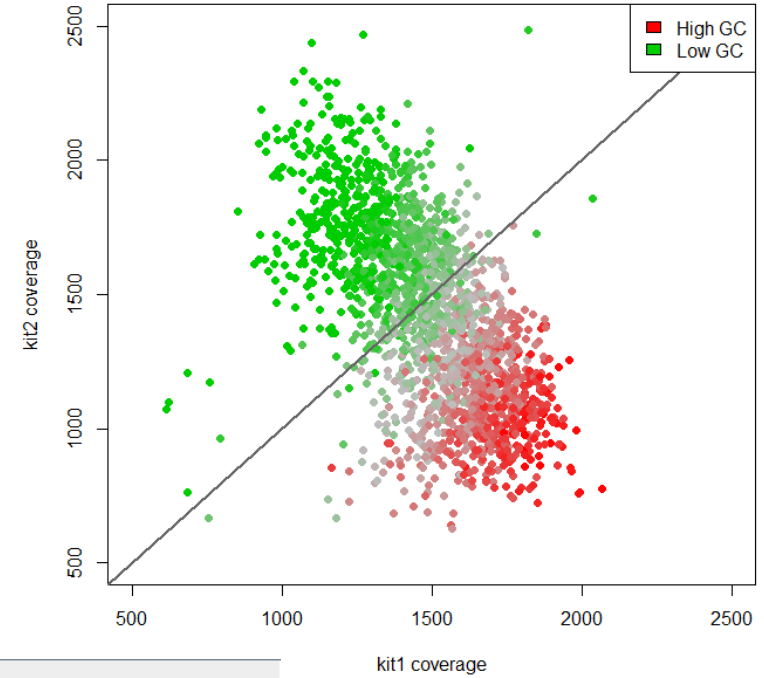
- Is a functional gene set enriched for genes in my hit list compared to a background set
- Are some genes **more likely** to turn up in the hits for technical reasons?
- Are some genes **never likely** to turn up in the hit list for technical reasons?

Biases

- All datasets contain biases
 - Technical
 - Biological
 - Statistical
- Biases can lead to incorrect conclusions
- We should be trying to spot these
 - Some are more obvious than others!

Technical Biases

- Simple GC bias from different polymerases in PCR



Statistical Biases

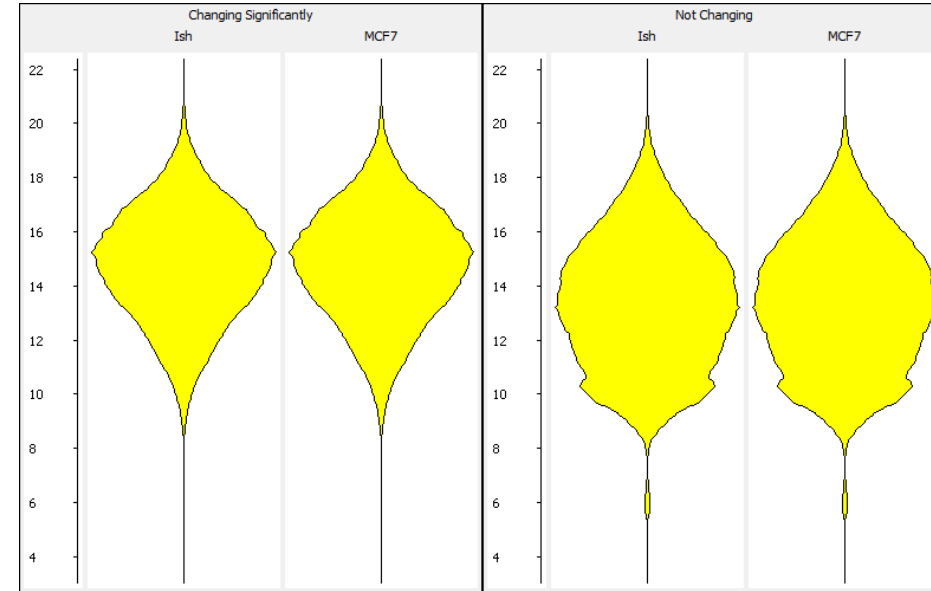
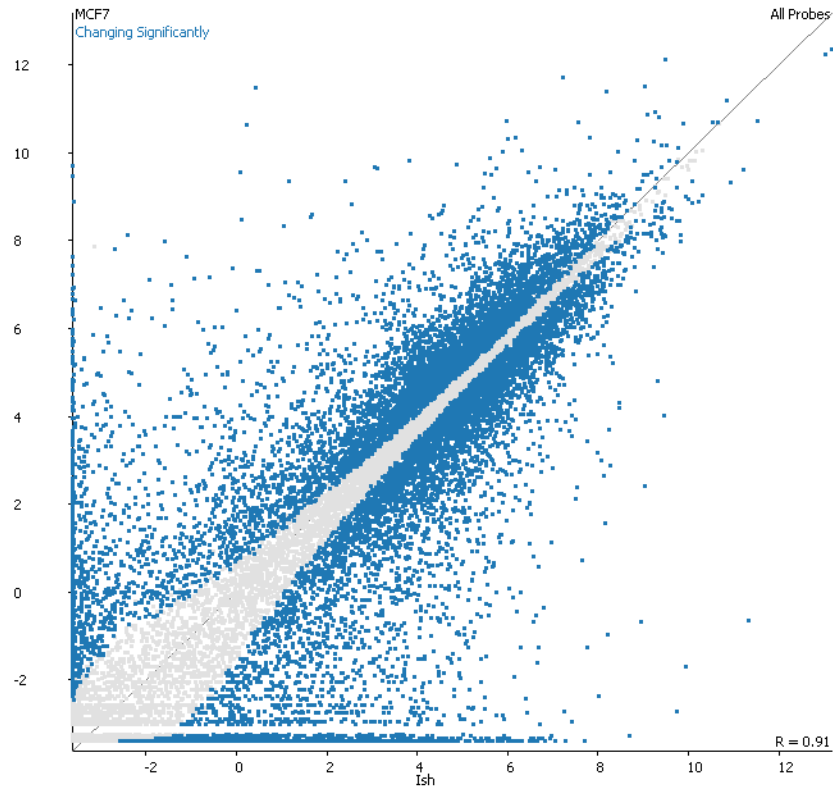
- The power to detect a significant effect is based on:
 - How big the change is
 - How well observed the data is (sample size)
- Lists of hits are often biased based on statistical power

RNA-Seq Statistical Biases

What determines whether a gene is identified as significantly differentially regulated?

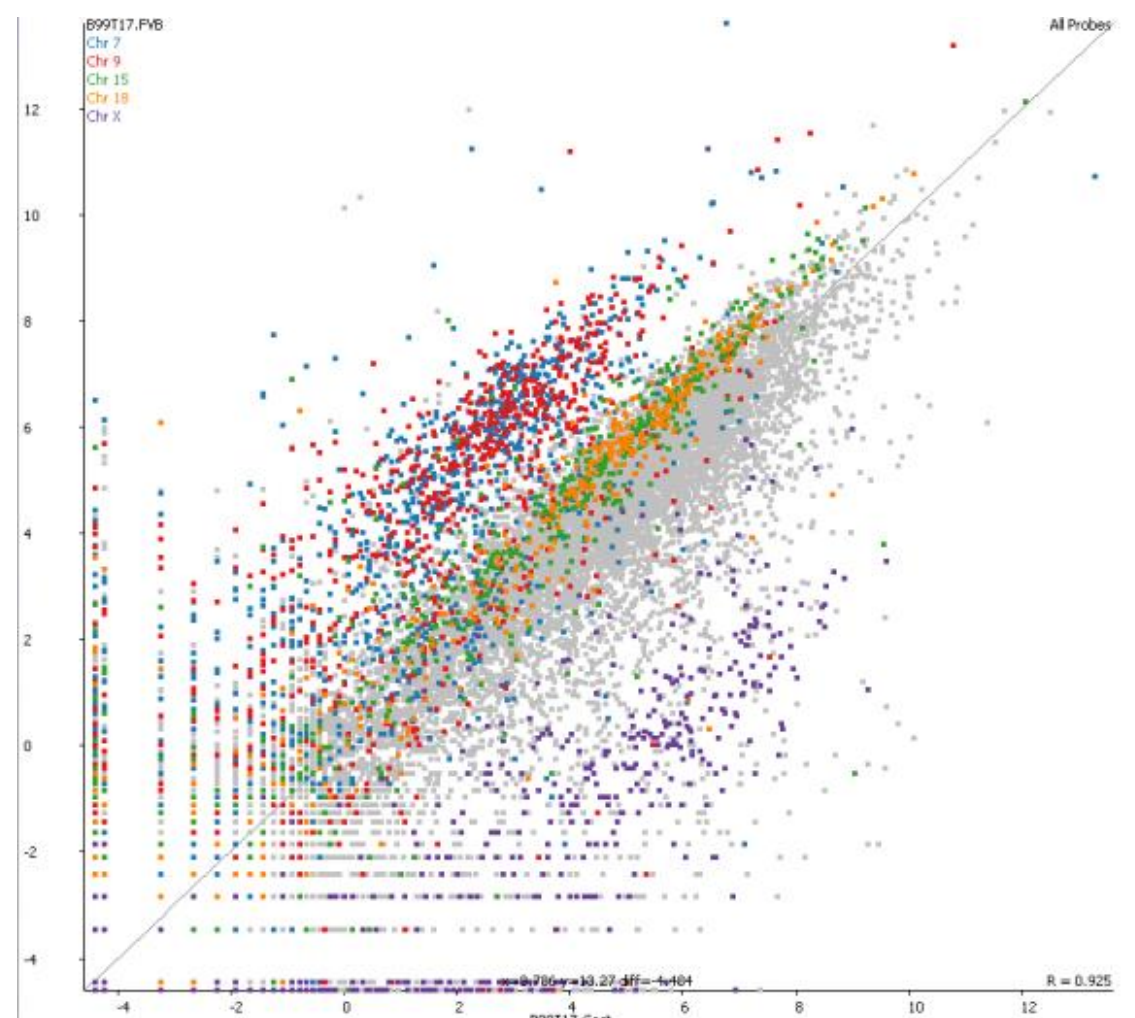
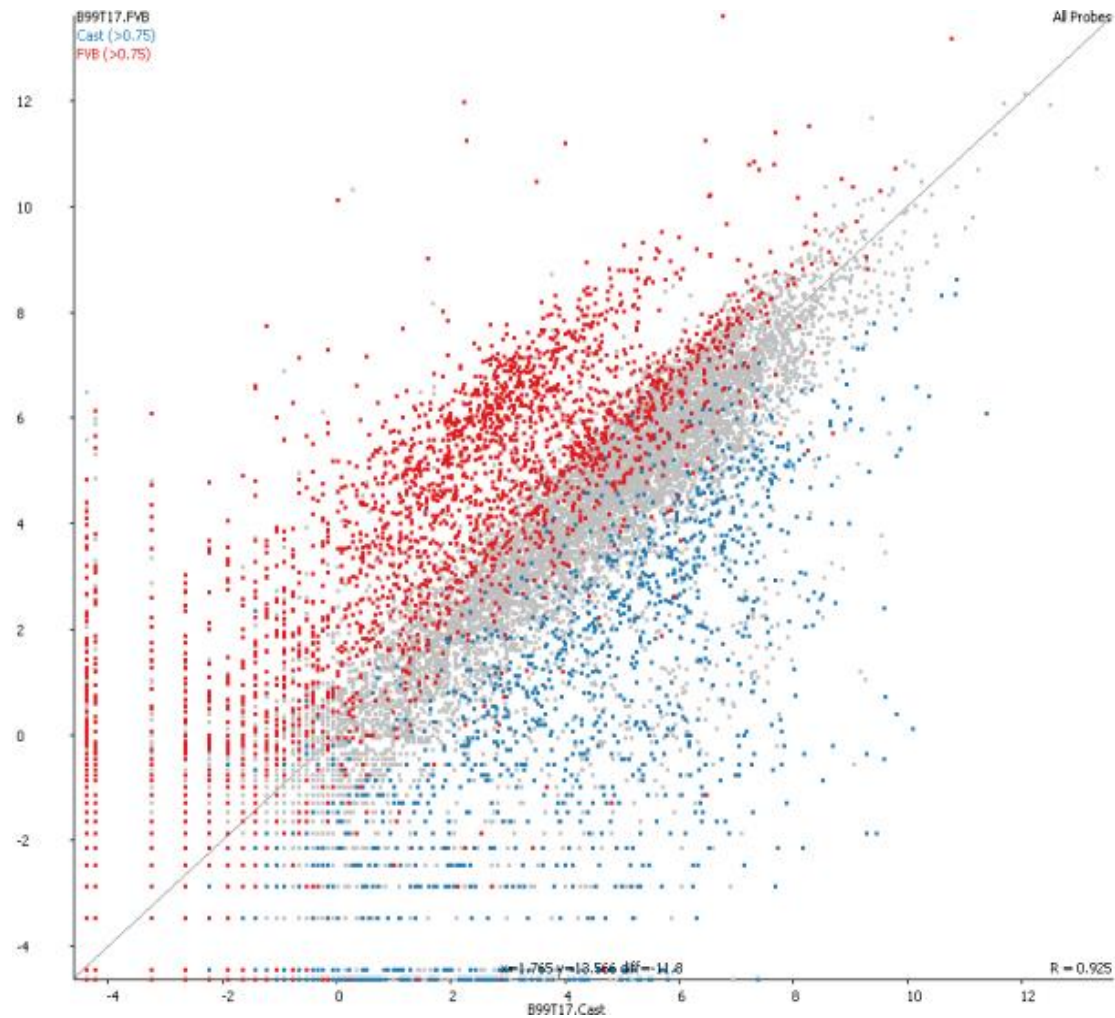
- The amount of change (fold change)
- The variability
- How well observed was it
 - How much sequencing was done overall?
 - How highly expressed was the gene?
 - How long was the gene?
 - How mappable was the gene?

RNA-Seq Statistical Biases



- Unlikely to ever see hits from genes which are
 - Lowly expressed
 - Short

Biological Biases



Biases Look Like Real Biology

Bias	Function	P-Value
High GC	DNA-Templated Transcription	2.00E-20
Low GC	GPCR Signalling	4.00E-12
Long Genes	Synapse	2.30E-30
Chr 18	Homophilic Cell Adhesion	1.01E-26

Research Article

Epigenetic Profiling of H3K4Me3 Reveals Herbal Medicine Jinfukang-Induced Epigenetic Alteration Is Involved in Anti-Lung Cancer Activity

Jun Lu,¹ Xiaoli Zhang,¹ Tingting Shen,¹ Chao Ma,² Jun Wu,¹ Hualei Kong,¹ Jing Tian,³ Zhifeng Shao,¹ Xiaodong Zhao,^{1,2} and Ling Xu^{2,4}

¹Shanghai Center for Systems Biomedicine, School of Biomedical Engineering, State Key Laboratory on Oncogene and Bio-ID Center, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

²Tumor Institute of Traditional Chinese Medicine, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, 725 South Wanping Road, Shanghai 200032, China

³College of Life Science, Northwest University, 229 Taibai Road, Xi'an 710069, China

Gene Ontology analysis indicates that these genes are involved in tumor-related pathways, including pathway in cancer, basal cell carcinoma, apoptosis, induction of programmed cell death, regulation of transcription (DNA-templated), intracellular signal transduction, and regulation of peptidase activity.

Traditional Chinese medicine Jinfukang (JFK) has been clinically used for treating lung cancer. To examine whether epigenetic modifications are involved in its anticancer activity, we performed a global profiling analysis of H3K4Me3, an epigenomic marker associated with active gene expression, in JFK-treated lung cancer cells. We identified 11,670 genes with significantly altered status of H3K4Me3 modification following JFK treatment ($P < 0.05$). Gene Ontology analysis indicates that these genes are involved in tumor-related pathways, including pathway in cancer, basal cell carcinoma, apoptosis, induction of programmed cell death, regulation of transcription (DNA-templated), intracellular signal transduction, and regulation of peptidase activity. In particular, we found that the levels of H3K4Me3 at the promoters of *SUSD2*, *CCND2*, *BCL2A1*, and *TMEM158* are significantly altered in A549, NCI-H1975, NCI-H1650, and NCI-H2228 cells, when treated with JFK. Collectively, these findings provide the first evidence that the anticancer activity of JFK involves modulation of histone modification at many cancer-related gene loci.

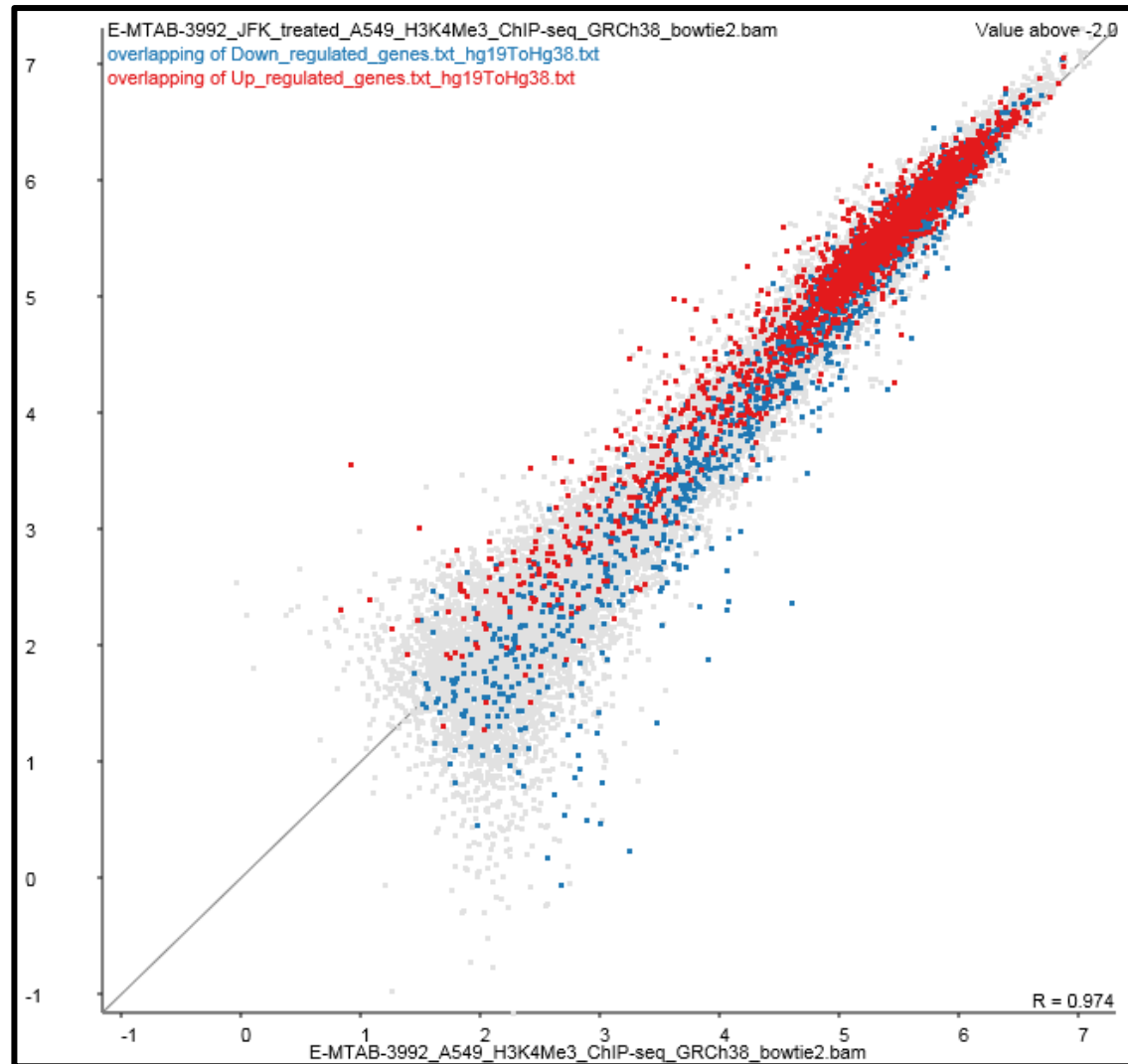
1. Introduction

Chromatin is the macromolecular complex of DNA and histone proteins that provides the scaffold for packaging the eukaryotic genome [1, 2]. Histones H2A, H2B, H3, and H4 are the basic components of nucleosomes, which form the fundamental unit of chromatin [3, 4]. Chemical modifications to the histones alter chromatin structure and regulate gene expression by altering noncovalent interactions within and between nucleosomes [2, 5]. H3K4Me3 is an active histone modification which is positively associated with gene expression [3, 6]. Previous studies have shown that the levels of H3K4Me3 modification are closely associated with the development, treatment, and diagnosis of

disease [7–9]. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) has been developed to systematically characterize the contribution of epigenetic regulation in various biological processes via genome-wide profiling of various chemical modifications of histone proteins and genomic DNA methylation [10].

Lung cancer has become the leading cause of cancer-related deaths worldwide [11]. Overall, only 16.8% of patients with lung cancer survive five years after their first definite diagnosis, mainly as a consequence of uncontrollable cell proliferation or tumor metastasis [12, 13]. Although various therapeutic interventions, including surgery, chemotherapy, and radiotherapy, have been developed to prolong the survival time of patients, drug side effects, pain, and emaciation

Bias or Biology?



ChIP



Input

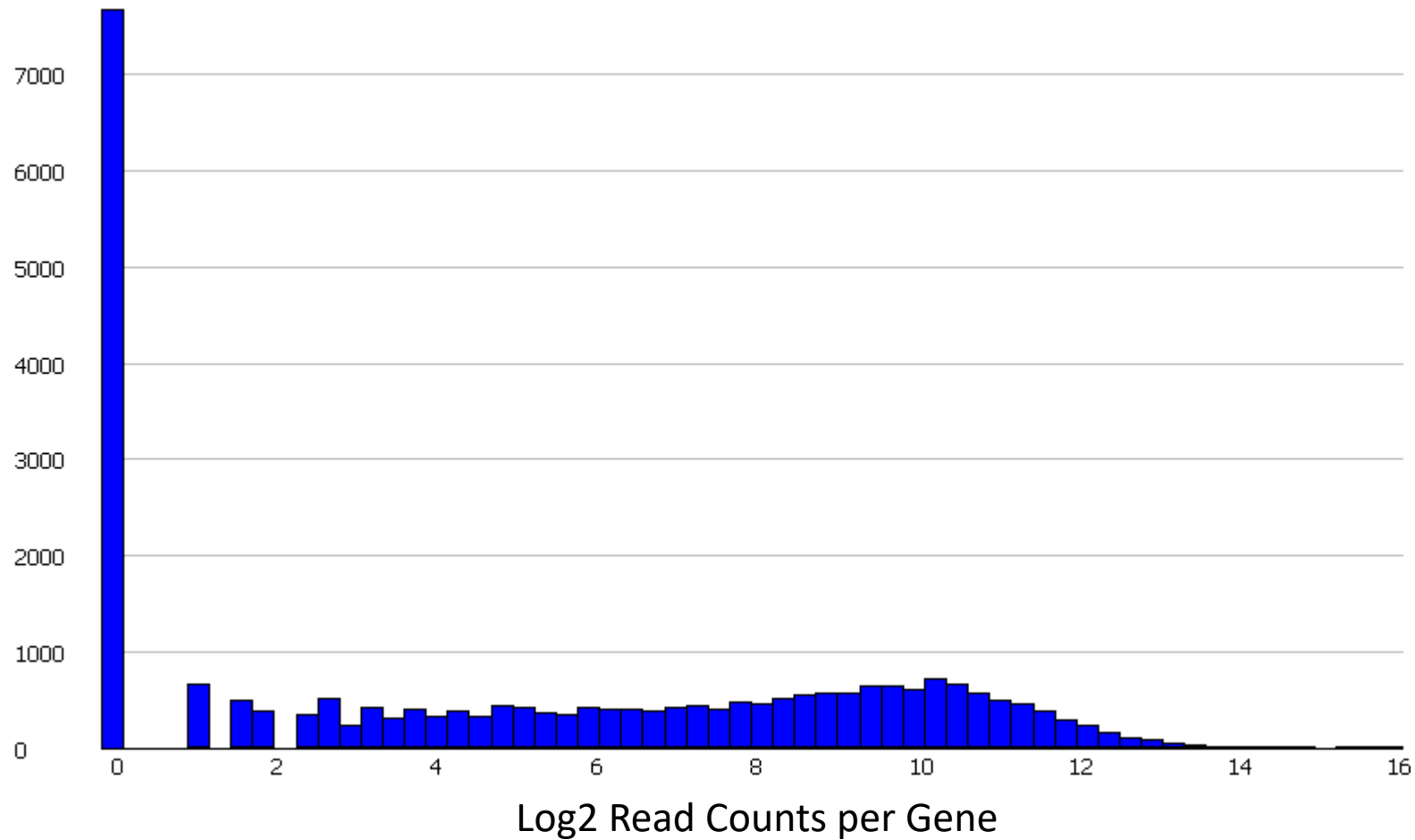
What can you do?

- Think about whether you're likely to have expected biases in your experiment.
- Look for unexpected biases.
 - Sometimes the bias *is* the interesting biology
- Use custom backgrounds during Gene Set Analysis to help minimise bias (if a tool supports it)

Using a background list can make a huge difference

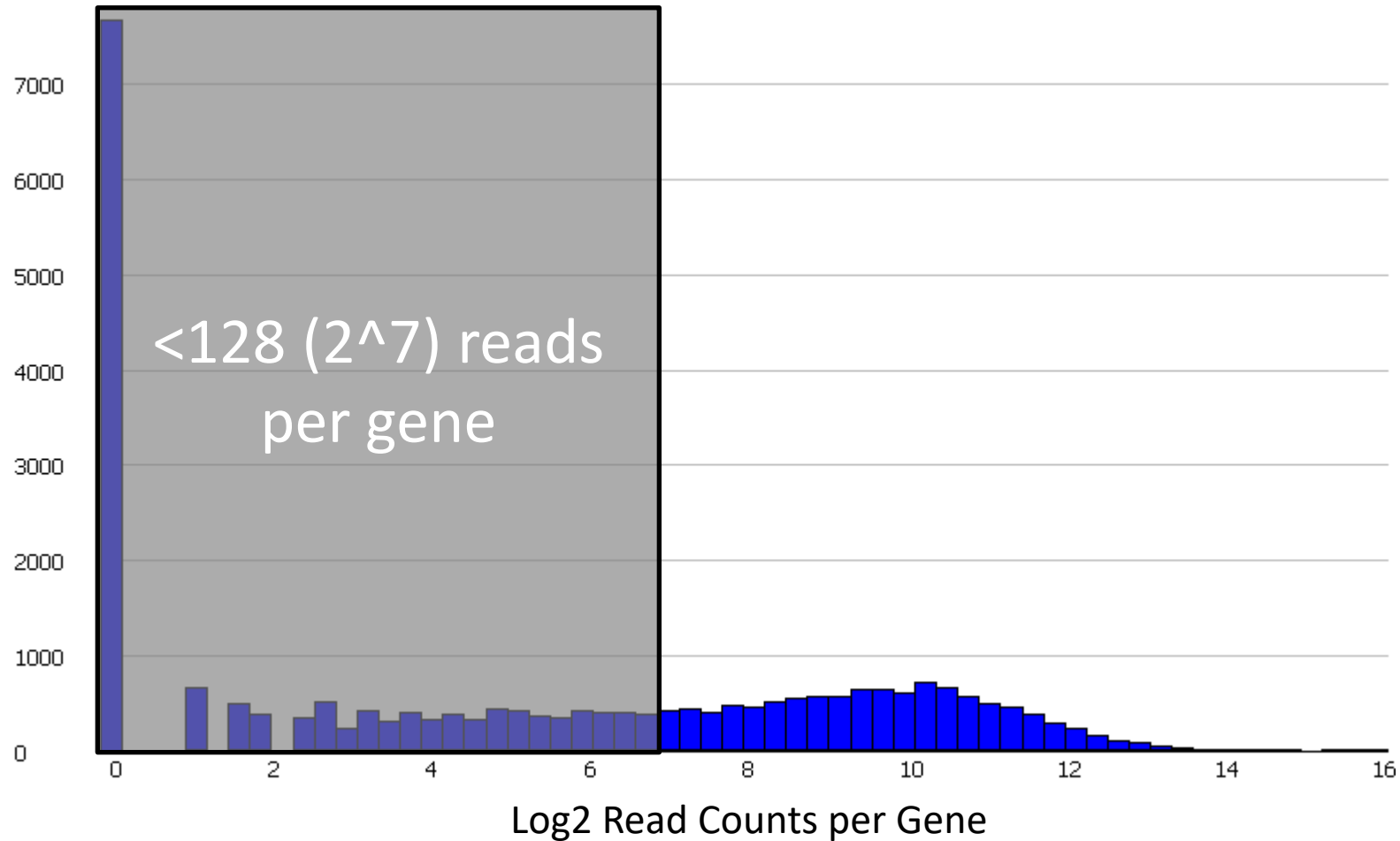
- What genes were you likely to see?
 - Some are technically impossible
 - Membrane proteins in LC-MS
 - Small-RNA in RNA-Seq
 - Some are much less likely
 - Unexpressed or low expressed in RNA-Seq
 - Unmappable in ChIP-Seq
 - Low CpG content in BS-Seq
- Make a list of what you ***could*** have seen, and set that as the background.

Expressed Genes



26,127 Genes Measured

Expressed Genes

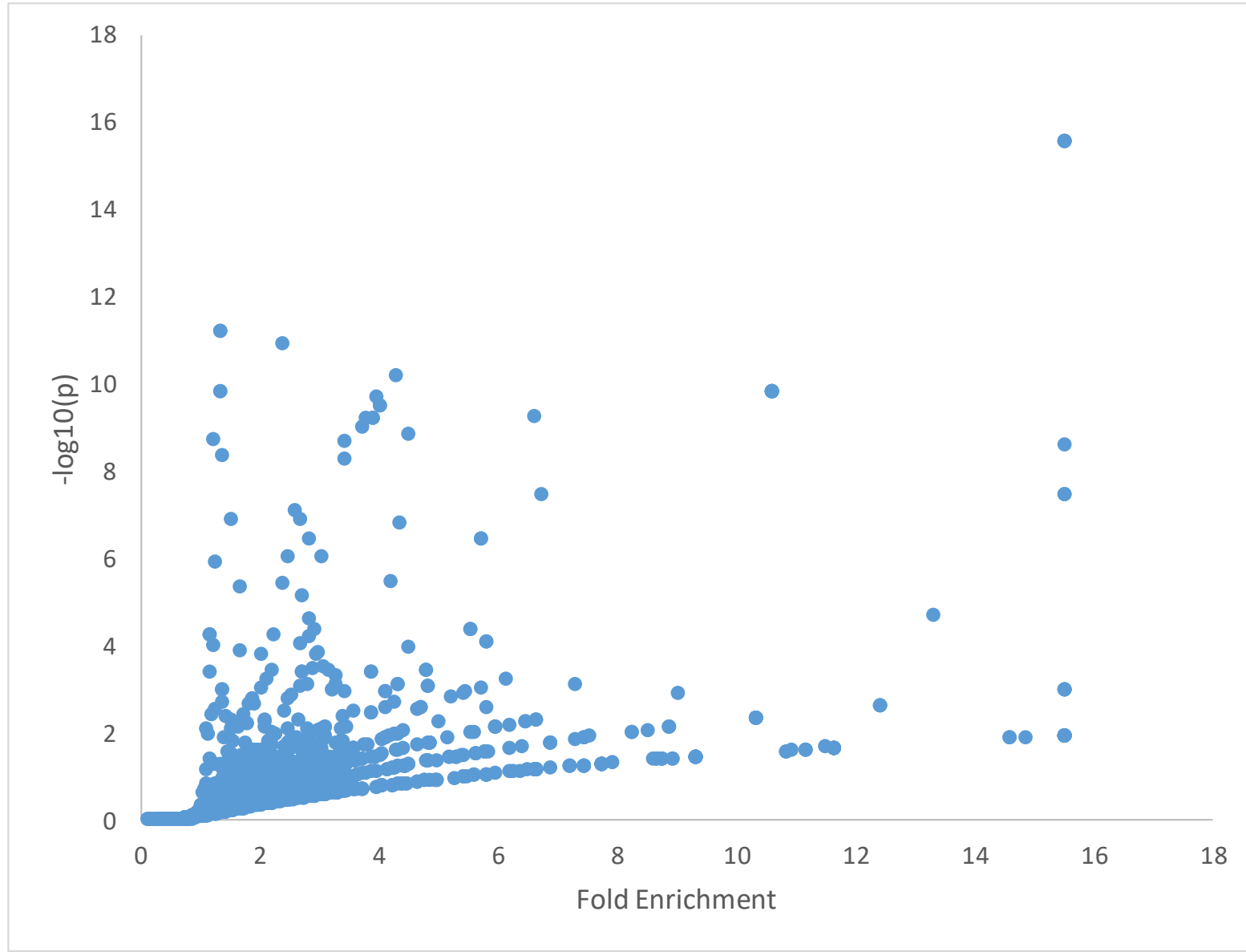


10,378 Genes Realistically Measured (40%)

Statistical biases affect gene sets too

- Fisher's test is powered by
 - Magnitude of change
 - Observation level
- Big lists have more power to detect change
- Small lists are very difficult to detect
- Some tools allow you to exclude the largest gene set categories. We often use categories with between 50 – 500 genes in to get power and specificity
- Always look at the enrichment and the p-value when deciding what is interesting

Fold Change and p-value



Other biases: Random Genomic Positions

- Relating genomic positions to genes
- Find closest gene
 - Synapse, Cell Junction, postsynaptic membrane ($p=8.9e-12$)
 - Membrane ($p=4.3e-13$)
- Find overlapping genes
 - Plekstrin homology domain ($p=1.8e-7$)
 - Ion transport ($p=7.1e-7$)

Creating a background list with the same biases as your hit list will alleviate the artefacts.

Stuff which turns up more than it should...

- Did a trawl through GEO RNA-Seq datasets
 - Downloaded pairs of samples which are supposed to be biological replicates
 - Found changing genes
 - Ran GO searches
- Many gene sets give hits. Some categories turn up very often
 - Ribosomal
 - Cytoskeleton
 - Extracellular
 - Secreted
 - Translation

Welcome to GOliath

Select species	<input type="text" value="Homo_Sapiens/Dec_18"/>
Min Category Size	<input type="text" value="50"/>
Max Category Size	<input type="text" value="500"/>
Gene List	Background List (optional)
<div>Paste Gene Names here</div>	<div>Paste Gene Names here</div>
<input type="text" value="Query name (optional)"/>	
<input type="button" value="Use example genes"/>	
<input type="button" value="Analyse my list"/>	

Results Table

Properties

Biases

Hit table

Copy

CSV

Excel

Print

Gene Set	Source	Query count	Background count	Category size	FDR	Enrichment	Potential bias
HALLMARK TNFA SIGNALING VIA NFKB	MSIGDB C2 HALLMARK TNFA SIGNALING VIA NFKB	17	200	200	1.384e-07	9.174	public_data
SIGNALING BY INTERLEUKINS	REACTOME R-HSA-449147.11	20	461	461	4.195e-05	4.683	high_transcripts
POSITIVE REGULATION OF CYTOKINE PRODUCTION	GOBP GO:0001819	16	355	355	0.0005164	4.865	public_data
HALLMARK IL2 STAT5 SIGNALING	MSIGDB C2 HALLMARK IL2 STAT5 SIGNALING	12	200	200	0.0009546	6.476	public_data
HALLMARK APOPTOSIS	MSIGDB C2 HALLMARK APOPTOSIS	10	160	160	0.003579	6.746	
APOPTOSIS	WIKIPATHWAYS 20190910 WP254 HOMO SAPIENS	8	87	87	0.003579	9.925	
REGULATION OF CYTOKINE SECRETION	GOBP GO:0050707	10	154	154	0.003579	7.009	
REGULATION OF INTERLEUKIN-6 PRODUCTION	GOBP GO:0032675	8	101	101	0.007389	8.549	
HALLMARK INFLAMMATORY RESPONSE	MSIGDB C2 HALLMARK INFLAMMATORY RESPONSE	10	200	200	0.008619	5.397	public_data
HALLMARK ALLOGRAFT REJECTION	MSIGDB C2 HALLMARK ALLOGRAFT REJECTION	10	200	200	0.008619	5.397	public_data

Search gene set

Search source

min query

min bg

min size

min FDR

min enrichment

Search bias

max query

max bg

max size

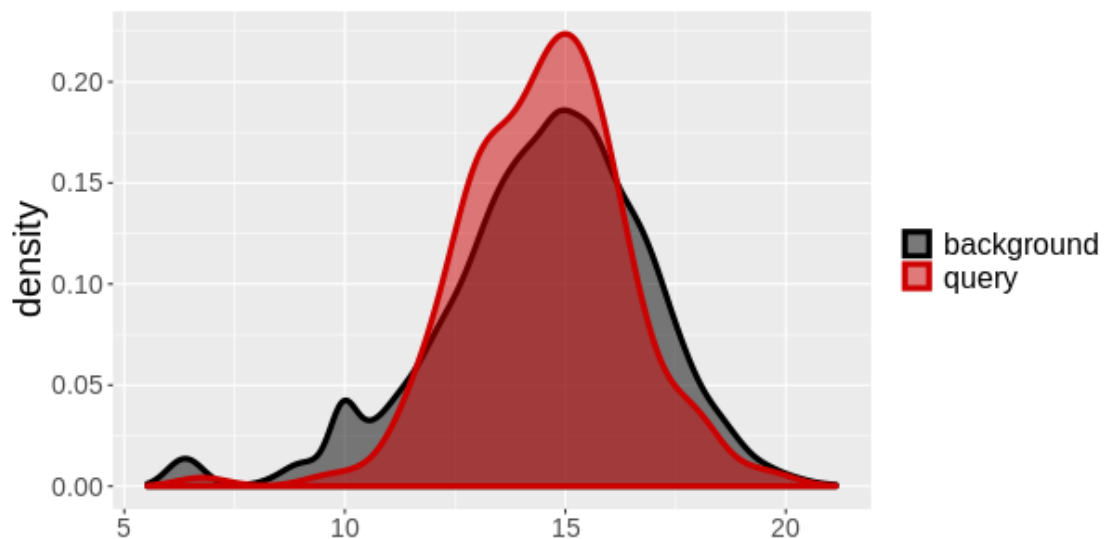
max FDR

max enrichment

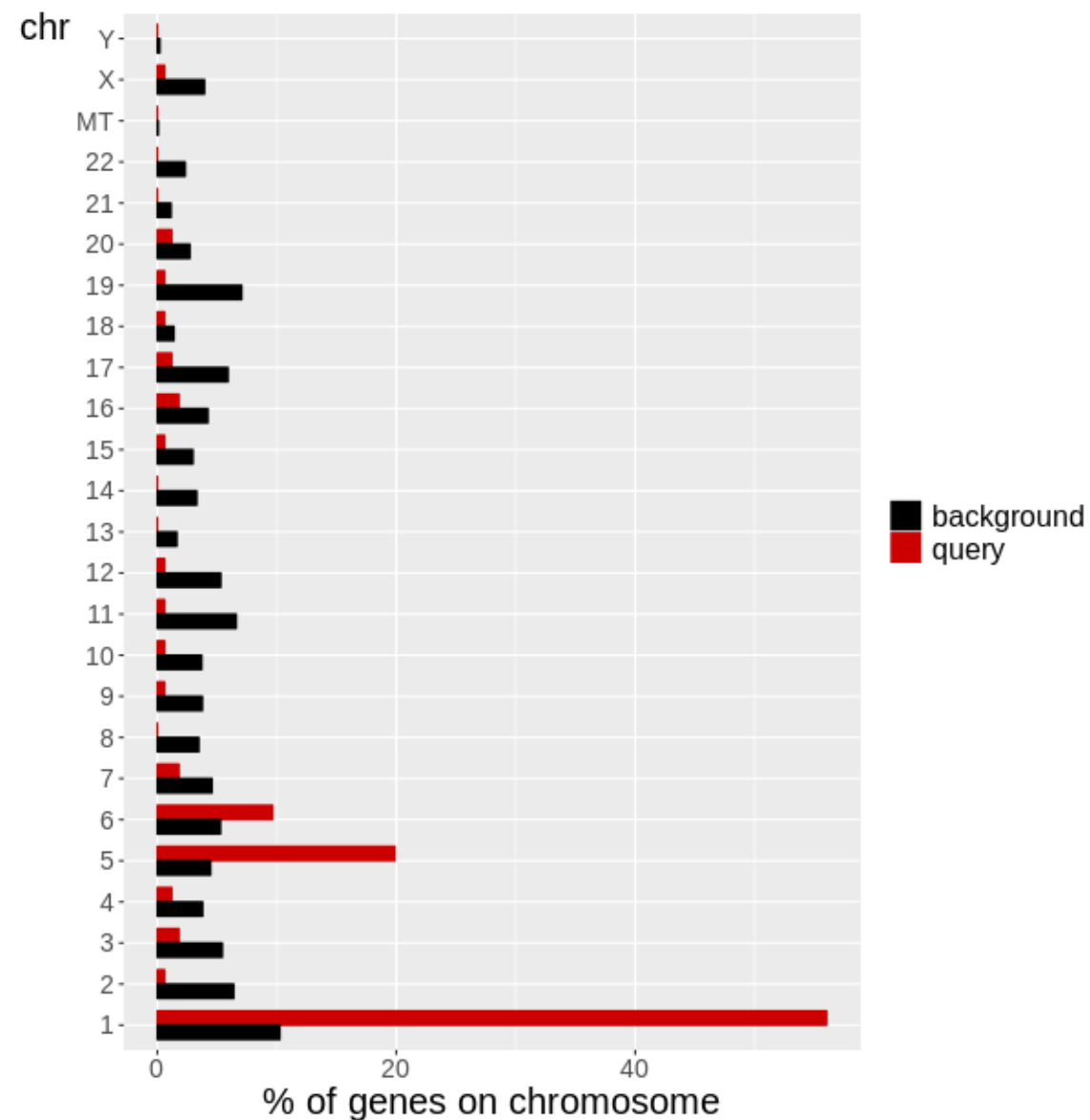
Checking for unexpected biases

- Do my hits look different from non-hits in factors which should be unrelated
 - Sequence composition
 - Genomic position
 - Gene Length
 - Number of splice variants
 - etc
- If a bias exists then is this the actual link between genes? If not then can I fix this by improving my background list?

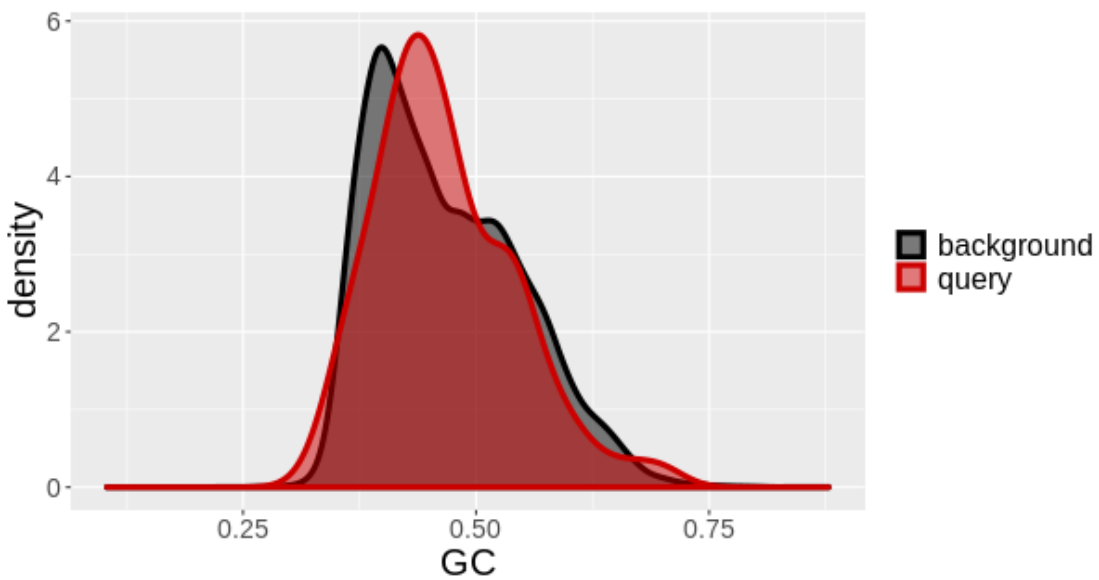
Gene lengths



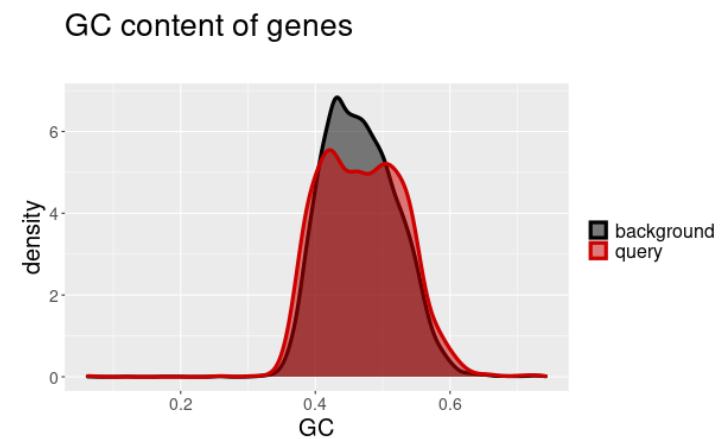
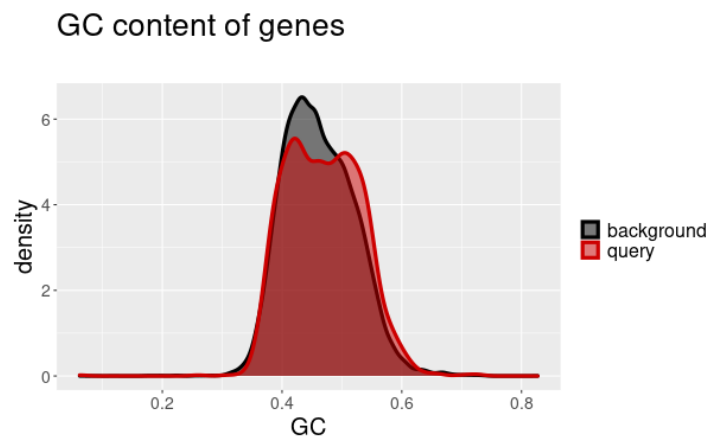
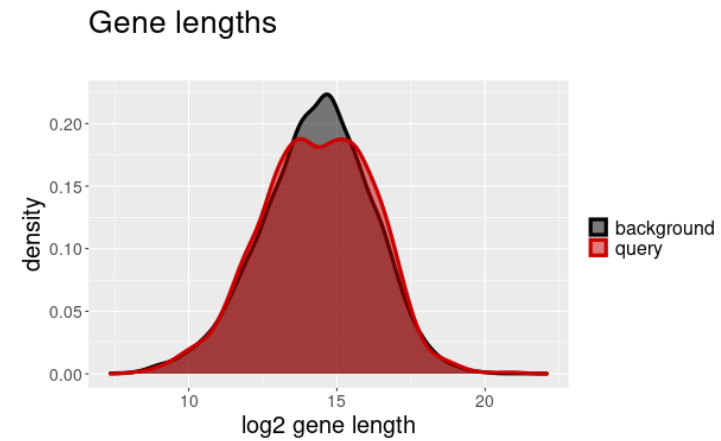
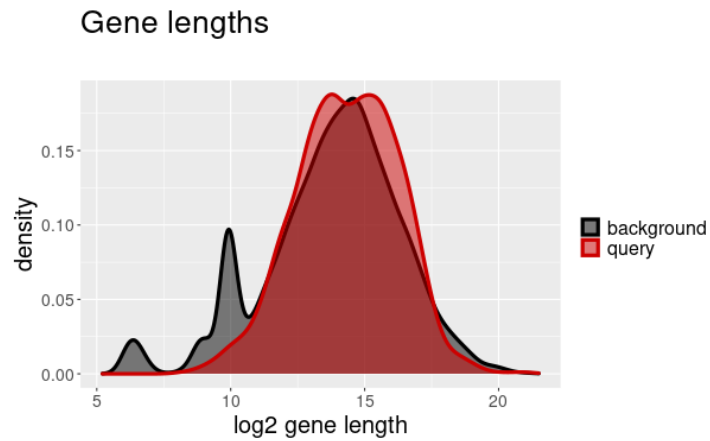
Chromosomal locations



GC content of genes



Custom backgrounds can make a difference



Custom backgrounds can make a difference

Top hits without correction

PLURINETWORK

POSITIVE REGULATION OF VASCULATURE DEVELOPMENT

POSITIVE REGULATION OF ANGIOGENESIS

HALLMARK E2F TARGETS

CHROMOSOME, CENTROMERIC REGION

DNA REPAIR

NEGATIVE REGULATION OF CELLULAR AMIDE METABOLISM

POSITIVE REGULATION OF ENDOTHELIAL CELL MIGRATION

NUCLEAR CHROMOSOME SEGREGATION

PID INTEGRIN1 PATHWAY

Top hits with correction

POSITIVE REGULATION OF VASCULATURE DEVELOPMENT

POSITIVE REGULATION OF ANGIOGENESIS

PID INTEGRIN1 PATHWAY

BETA1 INTEGRIN CELL SURFACE INTERACTIONS

INTEGRIN BINDING

ASSEMBLY OF COLLAGEN FIBRILS

NABA ECM REGULATORS

POSITIVE REGULATION OF ENDOTHELIAL CELL MIGRATION

RECEPTOR LIGAND ACTIVITY

STRIATED MUSCLE TISSUE DEVELOPMENT

Avoiding Biases

- Create a custom background if applicable
 - Should contain all genes which *could* have been in your hit list
 - May be a compromise, but it's better than nothing
 - Will limit which tools you can run
- Filter your tested gene sets
 - Remove large over powered sets, or sets which are too small to achieve significance (~50 to ~500 is generally about right)
 - Check the hit gene sets for matches to known problematic sets

Programmatic Gene List Analysis in R

Simon Andrews

simon.andrews@babraham.ac.uk

clusterProfiler

This is the **released** version of clusterProfiler; for the devel version, see [clusterProfiler](#).

A universal enrichment tool for interpreting omics data

- Categorical or quantitative enrichment
- Multi-species support

Table of contents

[Introduction](#)

Part I: Semantic Similarity
Analysis

- [1 Overview of semantic similarity analysis](#)
- [2 GO semantic similarity analysis](#)
- [3 DO semantic similarity analysis](#)
- [4 MeSH semantic similarity analysis](#)

Part II: Enrichment analysis

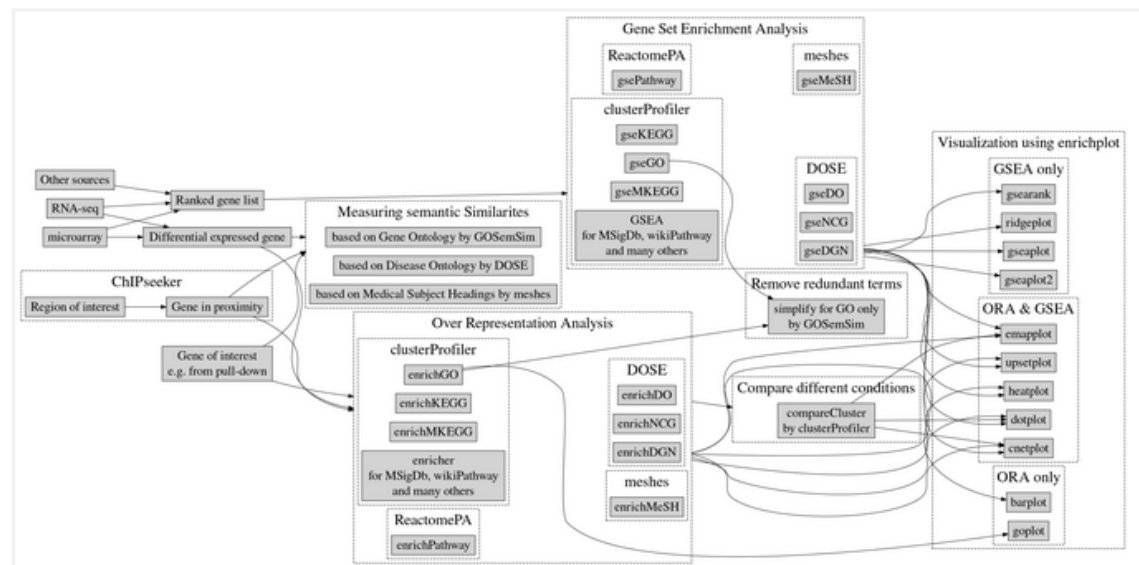
- [5 Overview of enrichment analysis](#)
- [6 GO enrichment analysis](#)
- [7 KEGG enrichment analysis](#)
- [8 WikiPathways analysis](#)
- [9 Reactome enrichment analysis](#)
- [10 Disease enrichment analysis](#)
- [11 MeSH enrichment analysis](#)
- [12 Universal enrichment analysis](#)
- [13 Functional enrichment analysis of genomic](#)



Introduction

ChIPseeker 1.30.3 clusterProfiler 4.3.1.900 DOSE 3.21.1.9002 enrichplot 1.15.3 GOSemSim 2.21.1
meshes 1.20.0

hello



Motivation

The book is meant as a guide for mining biological knowledge to elucidate or interpret molecular mechanisms using a suite of R packages, including [ChIPseeker](#), [clusterProfiler](#), [DOSE](#), [enrichplot](#), [GOSemSim](#), [meshes](#) and [ReactomePA](#). Hence, if you are starting to read this book, we assume you have a working knowledge of how to use R.

On this page

- [Introduction](#)
- [Motivation](#)
- [Citation](#)
- [Book structure](#)
- [Want to help?](#)

Genome Annotations

org.Hs.eg.db

This is the **released** version of org.Hs.eg.db; for the devel version, see [org.Hs.eg.db](#).

Genome wide annotation for Human

GO.db

This is the **released** version of GO.db; for the devel version, see [GO.db](#).

A set of annotation maps describing the entire Gene Ontology

org.Ag.eg.db
org.At.tair.db
org.Bt.eg.db
org.Cf.eg.db
org.Gg.eg.db
org.Pt.eg.db
org.Eck12.eg.db
org.Dm.eg.db
org.Hs.eg.db
org.Pf.plasmo.db
org.Mm.eg.db
org.Ss.eg.db
org.Rn.eg.db
org.Mmu.eg.db
org.Ce.eg.db
org.Xl.eg.db
org.Sc.sgd.db
org.Dr.eg.db

Anopheles
Arabidopsis
Bovine
Canine
Chicken
Chimp
E coli K12
Fly
Human
Malaria
Mouse
Pig
Rat
Rhesus
Worm
Xenopus
Yeast
Zebrafish

Installation and loading

```
install.packages("BiocManager")
```

```
BiocManager::install("clusterProfiler")
```

```
BiocManager::install("org.Hs.eg.db")
```

```
library(clusterProfiler)
```

```
library(org.Hs.eg.db)
```

```
library(tidyverse)
```

Different types of enrichment analysis

- Categorical enrichment
 - `enrichGO`
 - `enrichPathway`
- Quantitative enrichment
 - `gseGO`
 - `gsePathway`

Categorical Enrichment Analysis

Categorical Gene Ontology Enrichment

- Input requirements
 - Vector of hit gene ids
 - Keytype for the ids
 - Vector of all analysed gene ids

```
> keytypes(org.Hs.eg.db)
```

[1]	"ACCNUM"	"ALIAS"	"ENSEMBL"	"ENSEMBLPROT"	"ENSEMBLTRANS"
[6]	"ENTREZID"	"ENZYME"	"EVIDENCE"	"EVIDENCEALL"	"GENENAME"
[11]	"GENETYPE"	"GO"	"GOALL"	"IPI"	"MAP"
[16]	"OMIM"	"ONTOLOGY"	"ONTOLOGYALL"	"PATH"	"PFAM"
[21]	"PMID"	"PROSITE"	"REFSEQ"	"SYMBOL"	"UCSCKG"
[26]	"UNIPROT"				

Picking the correct keytype

```
select(  
  org.Hs.eg.db,  
  keys=keys(org.Hs.eg.db)[1:10],  
  columns=c("ENSEMBL", "SYMBOL", "GENENAME")  
)
```

ENTREZID	ENSEMBL	SYMBOL	GENENAME
1	ENSG00000121410	A1BG	alpha-1-B glycoprotein
2	ENSG00000175899	A2M	alpha-2-macroglobulin
3	ENSG00000291190	A2MP1	alpha-2-macroglobulin pseudogene 1
9	ENSG00000171428	NAT1	N-acetyltransferase 1
10	ENSG00000156006	NAT2	N-acetyltransferase 2
12	ENSG00000196136	SERPINA3	serpin family A member 3
13	ENSG00000114771	AADAC	arylacetamide deacetylase
14	ENSG00000127837	AAMP	angio associated migratory cell protein
15	ENSG00000129673	AANAT	aralkylamine N-acetyltransferase

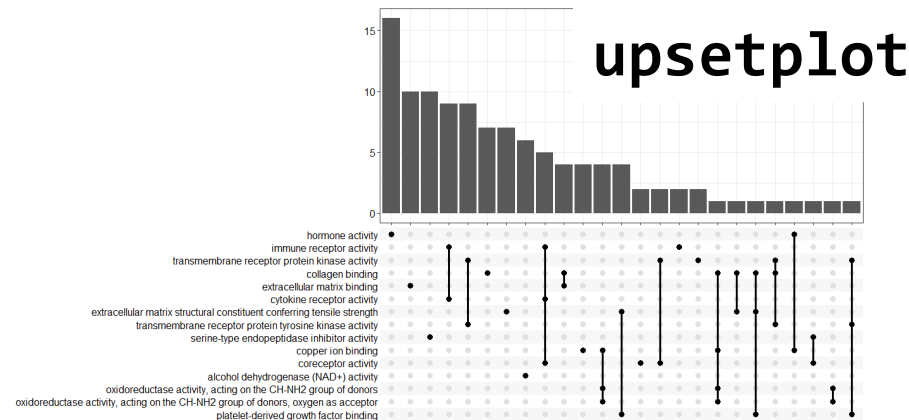
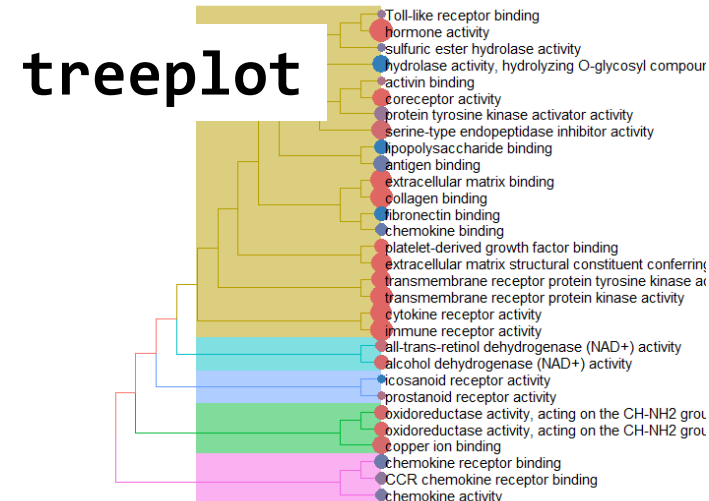
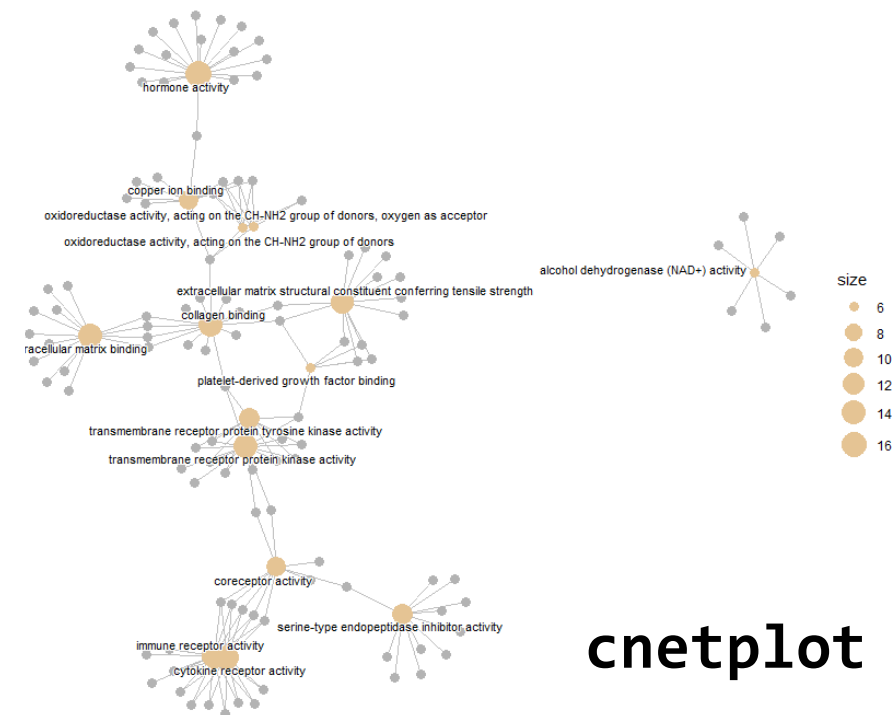
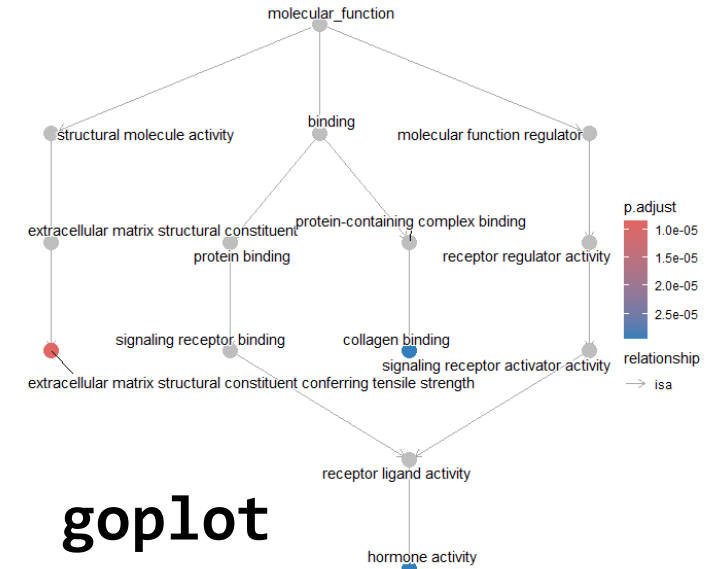
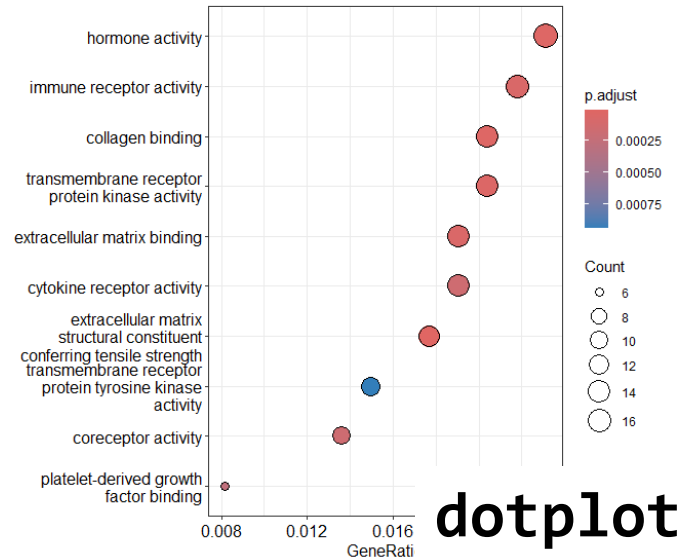
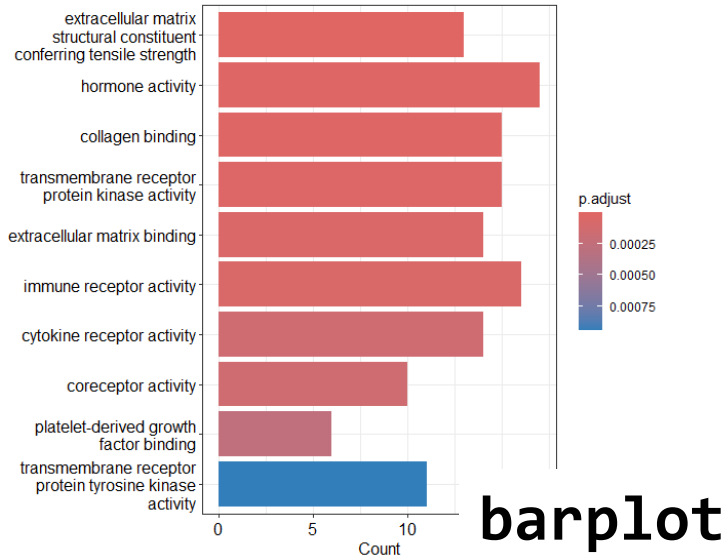
Running a categorical enrichment analysis

```
enrichGO(  
  up_genes,  
  OrgDb = org.Mm.eg.db,  
  keyType = "SYMBOL",  
  universe = background_genes,  
  ont = "BP",  
  minGSSize = 10,  
  maxGSSize = 100,  
  readable = TRUE  
) -> enrichgo_results
```

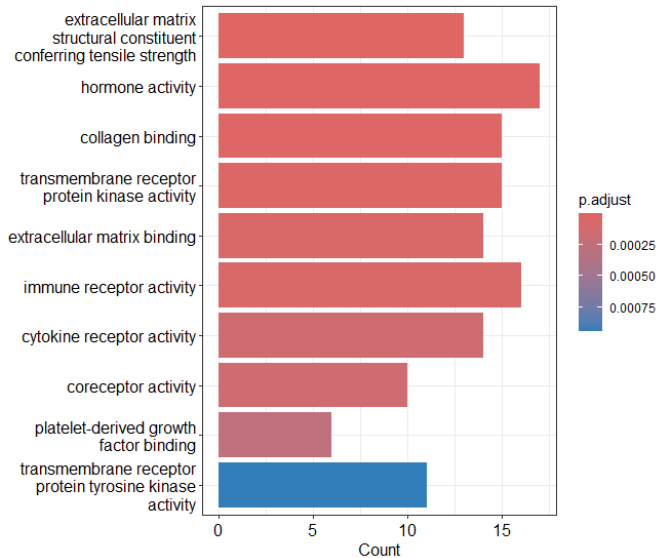
```
enrichgo_results@result |>  
  as_tibble()
```

ID <chr>	Description <chr>	GeneRatio <chr>	BgRatio <chr>	FoldEnrichment <dbl>	zScore <dbl>	pvalue <dbl>	p.adjust <dbl>	qvalue <dbl>
GO:0030199	collagen fibril organization	22/734	69/17422	7.567903	11.464270	3.666706e-14	1.249980e-10	1.017800e-10
GO:1903053	regulation of extracellular matrix organization	18/734	65/17422	6.572962	9.440344	1.102615e-10	1.879407e-07	1.530313e-07
GO:0072676	lymphocyte migration	19/734	81/17422	5.567632	8.641298	7.579679e-10	8.613041e-07	7.013197e-07
GO:0072678	T cell migration	15/734	56/17422	6.357775	8.421919	6.582620e-09	5.610038e-06	4.567992e-06
GO:1903487	regulation of lactation	7/734	10/17422	16.614986	10.358568	2.460222e-08	1.677380e-05	1.365812e-05
GO:2000401	regulation of lymphocyte migration	14/734	57/17422	5.829820	7.659729	7.008763e-08	3.549840e-05	2.890469e-05
GO:0031342	negative regulation of cell killing	10/734	27/17422	8.790998	8.496600	7.289200e-08	3.549840e-05	2.890469e-05
GO:2000404	regulation of T cell migration	12/734	42/17422	6.781627	7.867419	9.866697e-08	3.969711e-05	3.232350e-05
GO:0030168	platelet activation	18/734	97/17422	4.404562	7.051685	1.048032e-07	3.969711e-05	3.232350e-05
GO:0050830	defense response to Gram-positive bacterium	16/734	81/17422	4.688532	6.978169	2.173487e-07	7.409417e-05	6.033142e-05

Visualisation with enrichplot

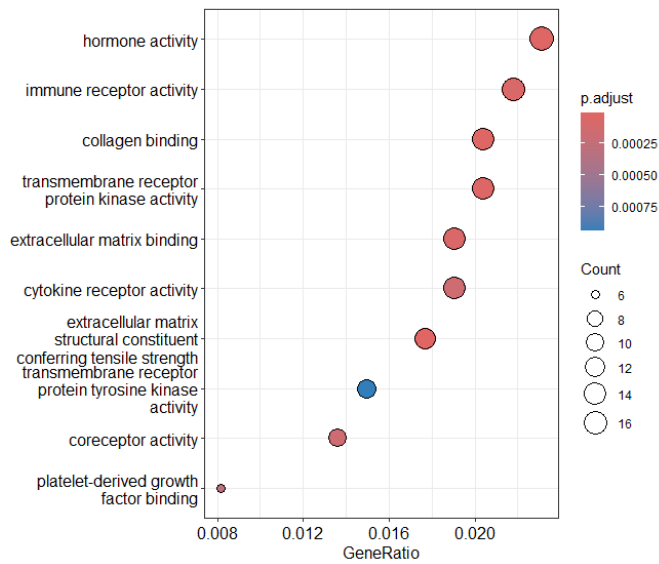


Visualisation with enrichplot



```
barplot(enrichgo_results, showCategory = 10)
dotplot(enrichgo_results, showCategory = 10)
```

```
enrichgo_results |>
  dotplot(
    x="FoldEnrichment",
    showCategory=20
  )
```



```
enrichgo_results |>
  dotplot(
    x="Count",
    showCategory=20
  )
```

GeneRatio

Fraction of background in hit list

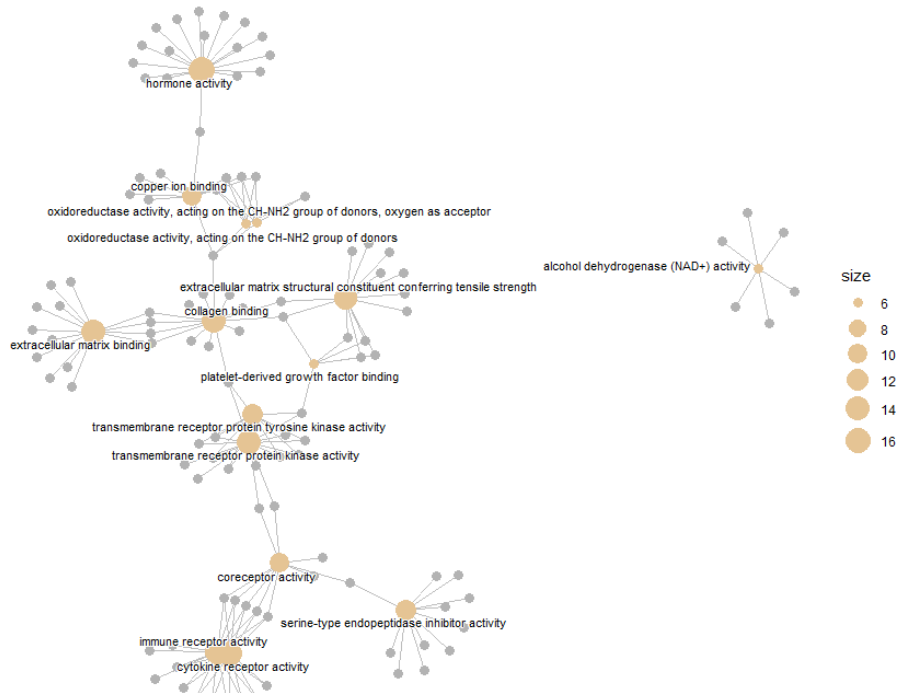
Count

Absolute number in hit list

Fold Enrichment

Fraction in hits over fraction in background

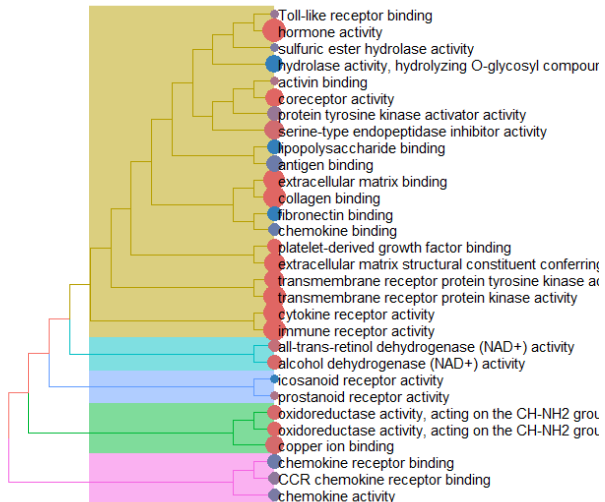
Visualisation with enrichplot



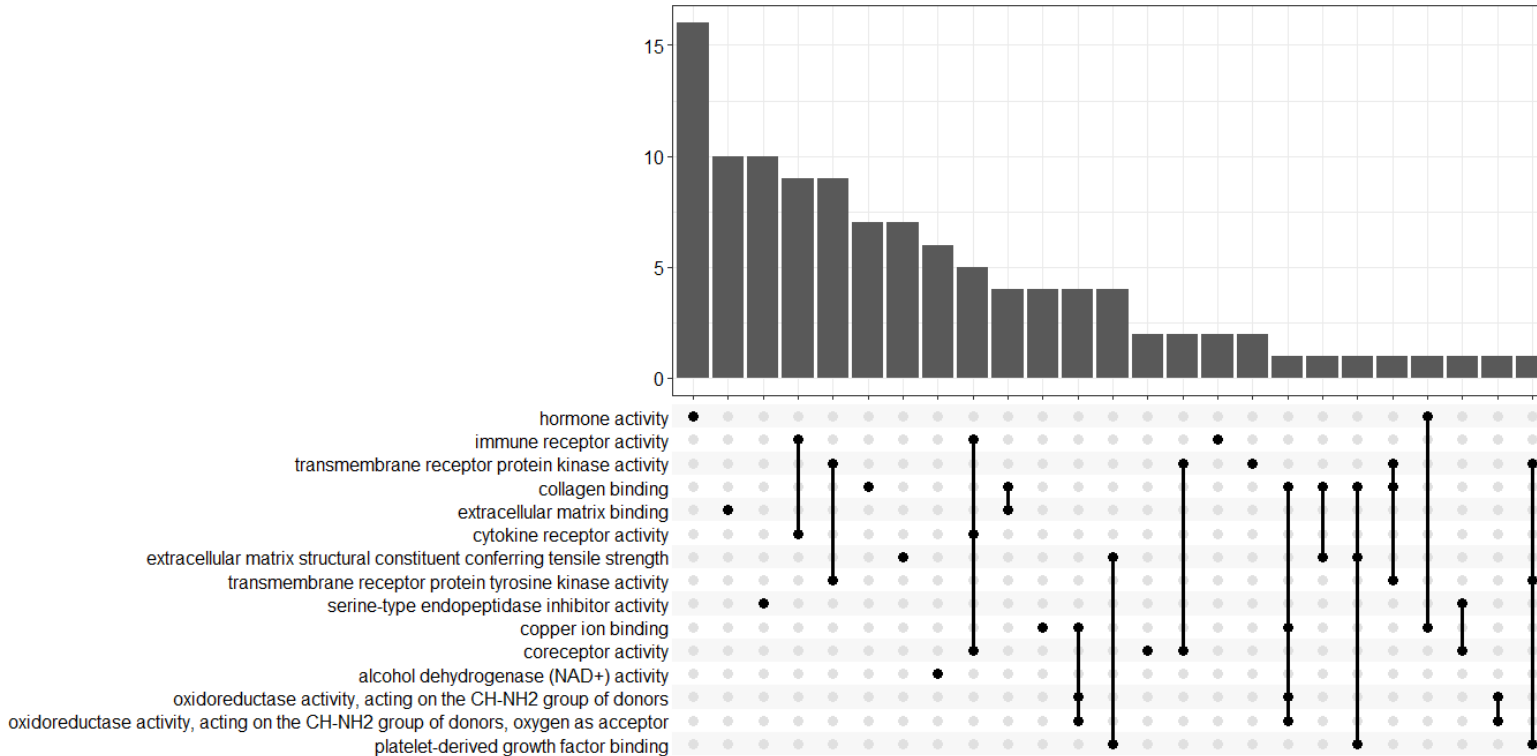
```
cnetplot(
  enrichgo_results,
  showCategory = 15,
  node_label="category"
)
```

- all
- none
- category
- gene

```
enrichgo_results |>
  pairwise_termsim() |>
  treeplot(
    showCategory=34,
    nCluster=7
  )
```



Visualisation with enrichplot

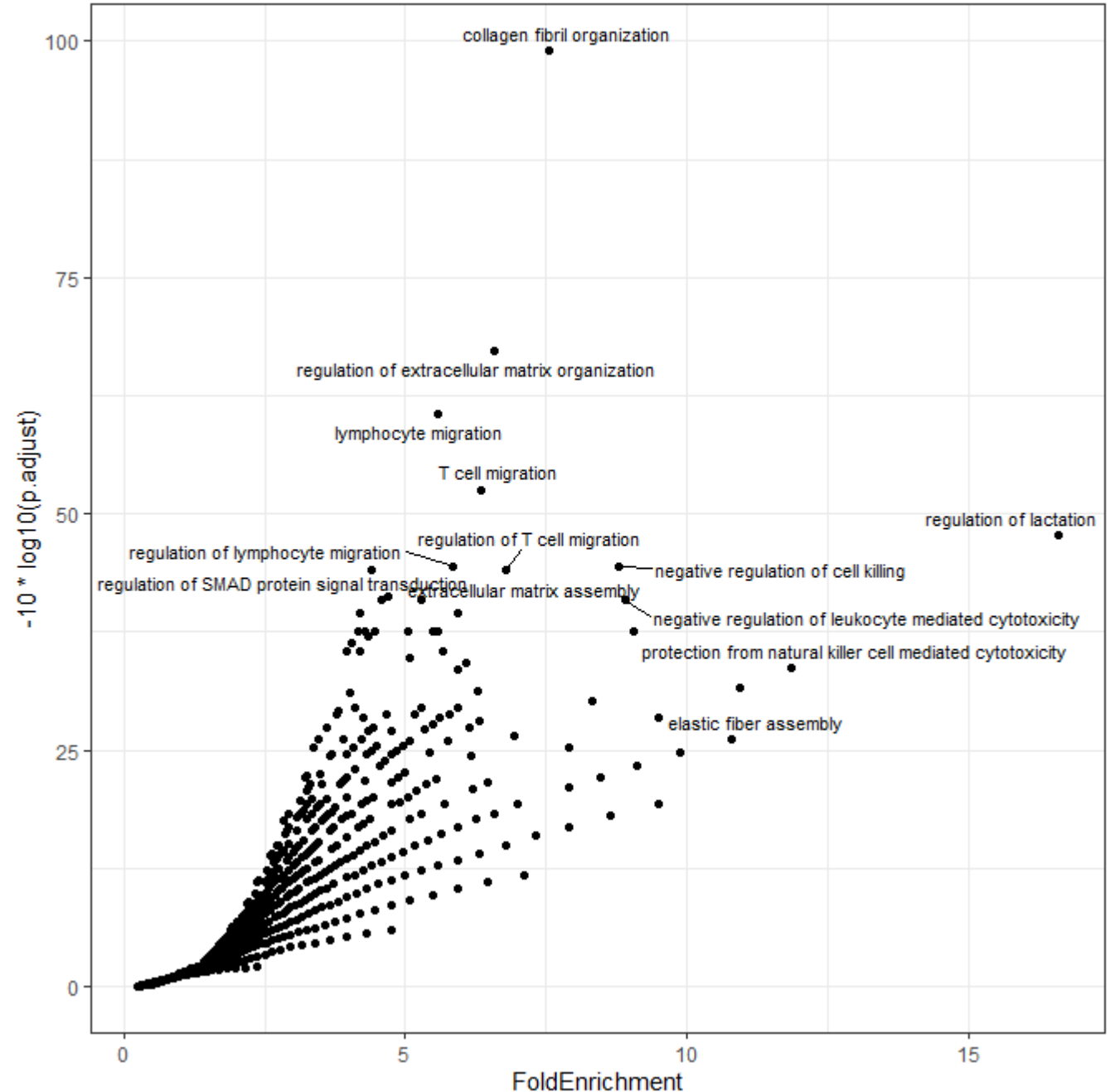


```
upsetplot(  
  enrichgo_results,  
  n=15  
)
```

Note: **not** showCategory in this plot

Custom Visualisation

```
enrichgo_results@result |>
  as_tibble(rownames="GOID") |>
  ggplot(aes(
    x=FoldEnrichment,
    y=-10*log10(p.adjust),
    label=Description
  )) +
  geom_point() +
  ggrepel::geom_text_repel(
    data = . %>% filter(
      FoldEnrichment > 5 &
      (-10*log10(p.adjust))>20
    ),
    size=3
  )
```



Categorical Enrichment with R Exercise

Quantitative Enrichment Analysis

Named vectors

```
c(1,5,3,6) -> my_data
```

```
my_data
```

```
[1] 1 5 3 6
```

```
c("simon","sarah","jo","hayley") -> names(my_data)
```

```
my_data
```

```
simon sarah jo hayley
```

```
1      5      3      6
```

GSEA input is a gene named vector of sorted (high to low) values

Running a quantitative enrichment analysis

Gene	Chromosome	Start	End	ID	Description	Type	P-value	FDR	ZScore
Xkr4	1	3276124	3741721	ENSMUSG000000051951	X-linked Kx blood group related 4 [Source:MGI Symbol;Acci...	gene	3.604699e-02	0.455440968	1.798524261
Rp1	1	4069780	4479464	ENSMUSG000000025900	retinitis pigmentosa 1 (human) [Source:MGI Symbol;Acc:MG...	gene	3.510687e-01	0.645575404	0.382436782
Sox17	1	4561154	4567577	ENSMUSG000000025902	SRY (sex determining region Y)-box 17 [Source:MGI Symbol;...	gene	8.803802e-02	0.587544382	-1.352936149
Mrpl15	1	4843429	4855962	ENSMUSG000000033845	mitochondrial ribosomal protein L15 [Source:MGI Symbol;A...	gene	3.934556e-01	0.645575404	0.270323783
Lyp1a1	1	4878011	4918633	ENSMUSG000000025903	lysophospholipase 1 [Source:MGI Symbol;Acc:MGI:1344588]	gene	3.866758e-01	0.645575404	0.287993759
Tcea1	1	4928037	4968132	ENSMUSG000000033813	transcription elongation factor A (SII) 1 [Source:MGI Symbol;...	gene	3.479390e-01	0.645575404	0.390890777
Rgs20	1	4979799	5140508	ENSMUSG000000002459	regulator of G-protein signaling 20 [Source:MGI Symbol;Acc...	gene	4.381771e-02	0.480191737	1.708004951
Atp6v1h	1	5140241	5232752	ENSMUSG000000033793	ATPase, H+ transporting, lysosomal V1 subunit H [Source:M...	gene	3.373898e-01	0.645575404	0.419597268
Oprk1	1	5658689	5676354	ENSMUSG000000025905	opioid receptor, kappa 1 [Source:MGI Symbol;Acc:MGI:97439]	gene	1.760258e-01	0.643769562	0.930617332
Npbwr1	1	5983926	5987617	ENSMUSG000000033774	neuropeptides B/W receptor 1 [Source:MGI Symbol;Acc:MGI...	gene	3.946832e-01	0.645575404	0.267133325

```
kat6ab_zscore |>
  arrange(desc(ZScore)) |>
  distinct(Gene, .keep_all = TRUE) -> data
```

```
data |>
  pull(ZScore) -> zscore_data
```

```
data$Gene -> names(zscore_data)
```

Fabp7	Gm3764	Reln	Gfap	Gm35040	C130071C03Rik
7.232035	6.434833	6.374436	6.281050	5.209227	5.187987

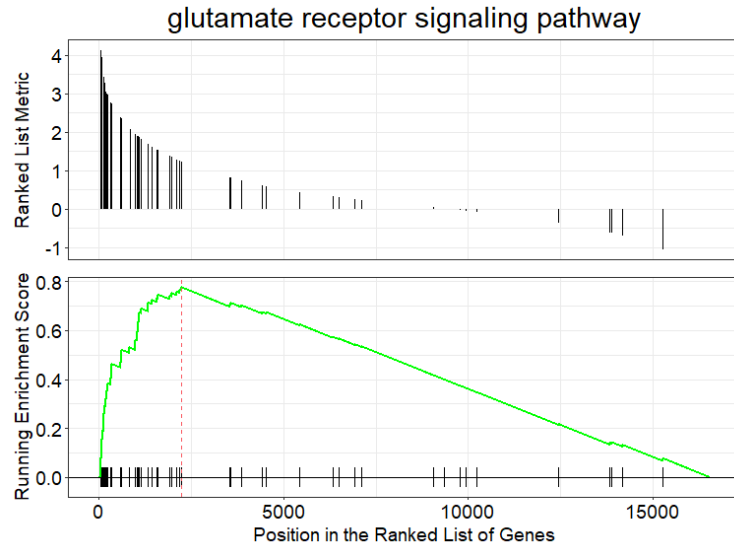
Running a quantitative enrichment analysis

```
gseGO(  
  zscore_data,  
  ont="BP",  
  OrgDb = "org.Mm.eg.db",  
  keyType = "SYMBOL",  
  minGSSize = 10,  
  maxGSSize = 100  
) -> gsego_result
```

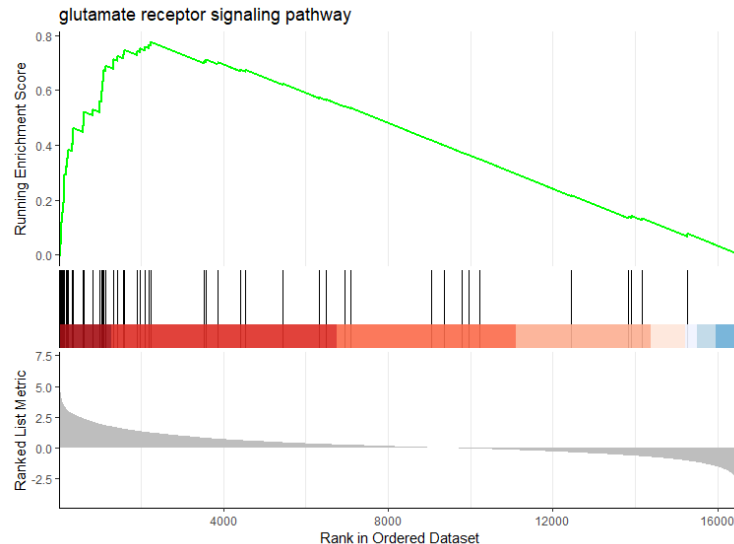
```
gsego_result@result |>  
  as_tibble()
```

ID <chr>	Description <chr>	setSize <int>	enrichmentScore <dbl>	NES <dbl>	pvalue <dbl>	p.adjust <dbl>
GO:0030199	collagen fibril organization	69	-0.6965226	-2.952459	1.000000e-10	4.995000e-08
GO:0007215	glutamate receptor signaling pathway	57	0.7761472	2.374072	1.000000e-10	4.995000e-08
GO:0099172	presynapse organization	66	0.7500618	2.348110	1.000000e-10	4.995000e-08
GO:0099505	regulation of presynaptic membrane potential	33	0.8316328	2.327325	1.000000e-10	4.995000e-08
GO:0099054	presynapse assembly	53	0.7530821	2.291861	1.000000e-10	4.995000e-08
GO:0051966	regulation of synaptic transmission, glutamatergic	83	0.7158265	2.287267	1.000000e-10	4.995000e-08
GO:0099174	regulation of presynapse organization	43	0.7731902	2.272698	1.000000e-10	4.995000e-08
GO:0019226	transmission of nerve impulse	100	0.6740946	2.209081	1.000000e-10	4.995000e-08
GO:1903539	protein localization to postsynaptic membrane	76	0.6860463	2.185224	1.000000e-10	4.995000e-08
GO:0062237	protein localization to postsynapse	79	0.6799646	2.170268	1.000000e-10	4.995000e-08

GSEA Visualisation



```
gseaplot(  
  gsego_result,  
  geneSetID = 2,  
  title = gsego_result$Description[2]  
)
```



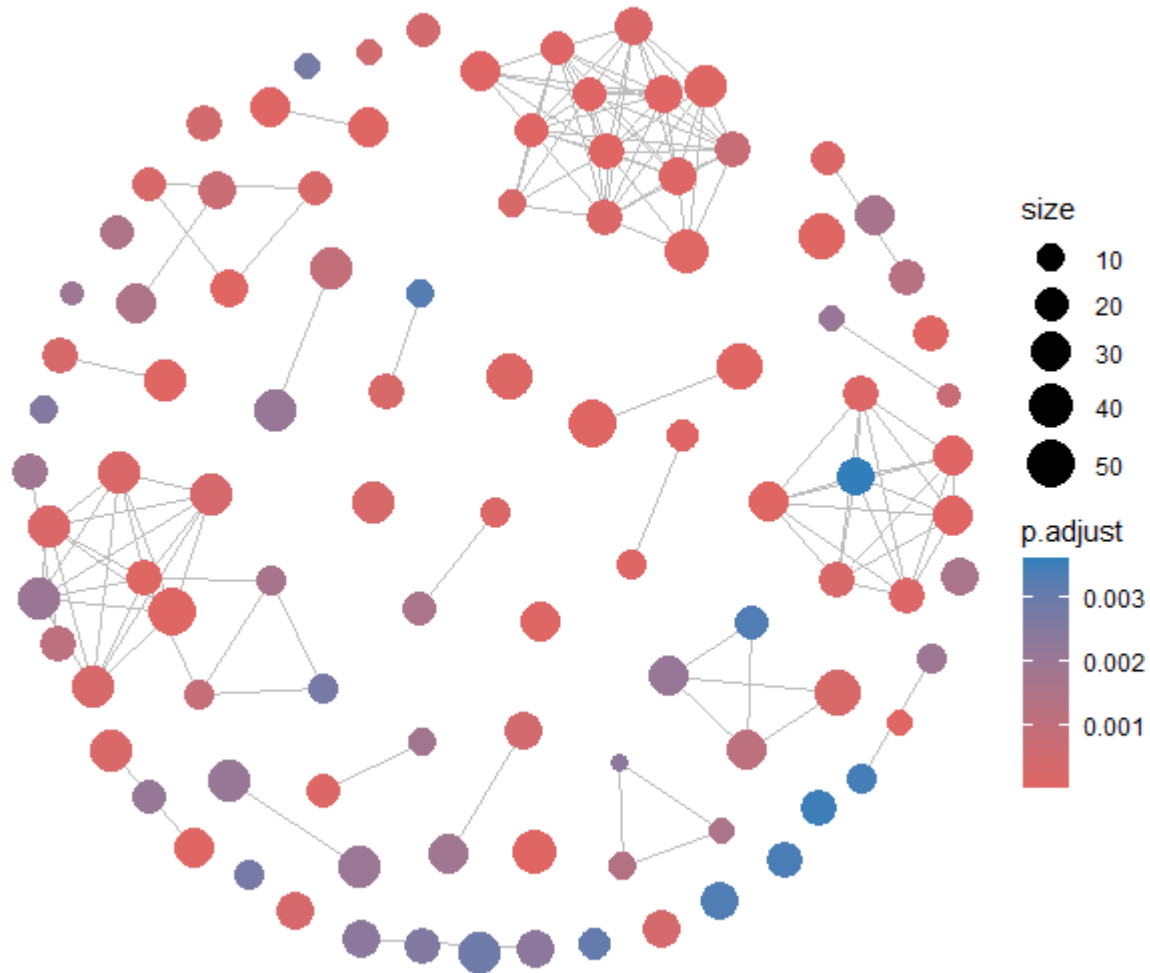
```
gseaplot2(  
  gsego_result,  
  geneSetID = 2,  
  title = gsego_result$Description[2]  
)
```

GSEA Visualisation



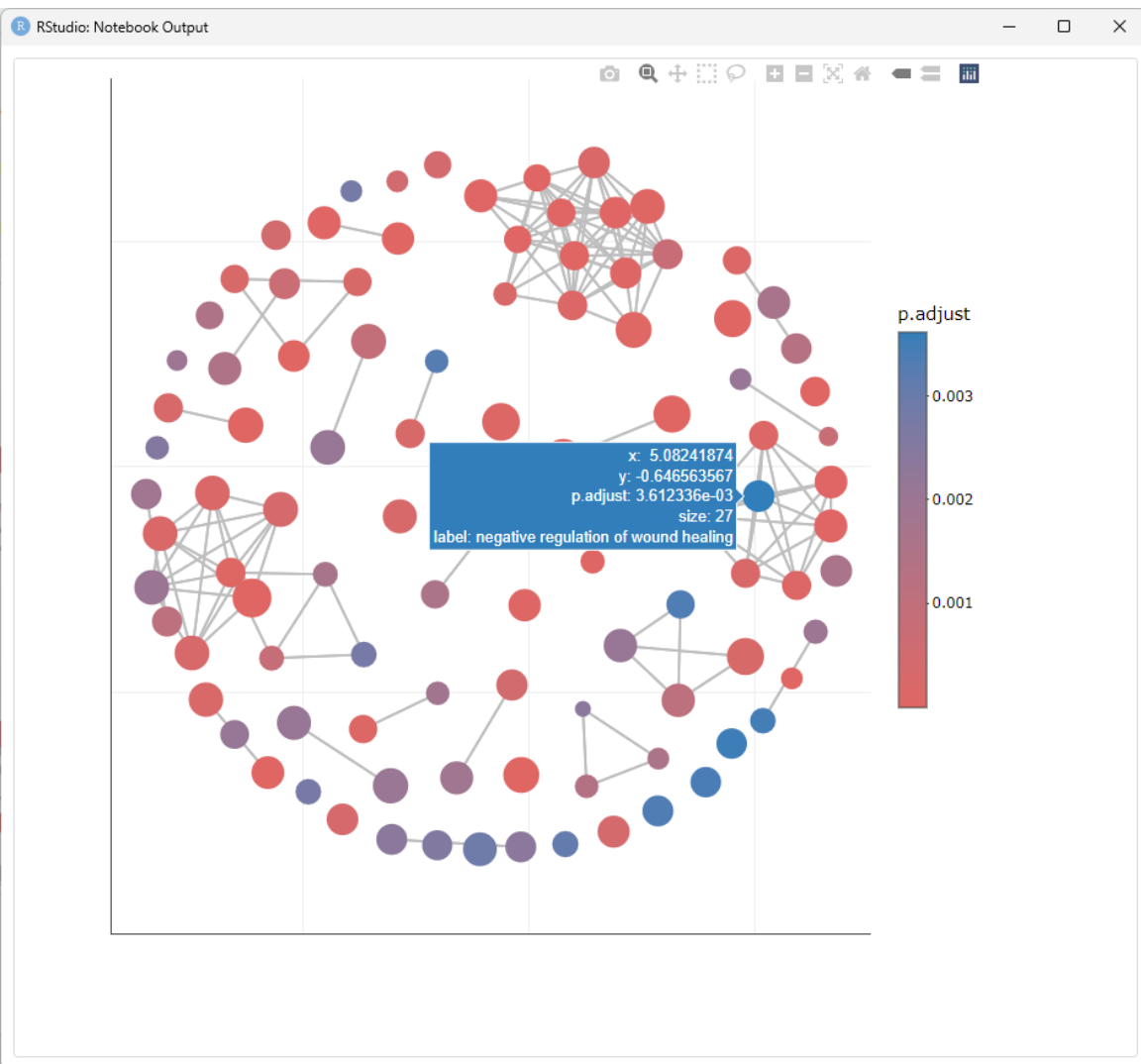
```
dotplot(  
  gsego_result,  
  x="NES",  
  showCategory=30  
)
```

GSEA Visualisation



```
gsego_result |>  
  filter(NES < 0) |>  
  pairwise_termsim() |>  
  emapplot(  
    min_edge=0.5,  
    showCategory=100,  
    node_label="none"  
  )
```

Interactive visualisation (kludge!)



```
gsego_result |>
  filter(NES < 0) |>
  pairwise_termsim() |>
  emapplot(min_edge=0.5,
    showCategory=100,
    node_label="category") -> temp
```

```
temp$layers[[2]]$mapping$label =
temp$data$label
```

```
plotly::ggplotly(temp, width = 800, height=800)
```

Quantitative Enrichment with R Exercise