

Extracting Biological Information from Gene Lists

Simon Andrews, Laura Biggins, Boo Virk

simon.andrews@babraham.ac.uk

laura.biggins@babraham.ac.uk





















v2023-01



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]
Ermn	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, ...

[Biaqi, 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-1) [3].
- On the other hand, gp34 was not expressed on these cells, although its expression is also known to be associated with [HTLV-1-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].



Functional analysis relates hits to existing knowledge

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

Most functional analysis starts from gene lists

- Many considerations
 - Other start points
 - Genomic positions
 - Transcripts / Proteins
 - Gene nomenclature
 - Annotation sources / versions
- Types of list
 - Categorical (hit or not a hit)
 - Ordered
 - Quantitative

A functional gene set provides a group of genes with a common biological relationship

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	Organism	PANTHER family	Type	Source	Synonyms
Hoxc4	homeobox C4	Mus musculus	family not named pthr24326	protein	MGI	Hox-3.5
Ing2	inhibitor of growth family, member 2	Mus musculus	inhibitor of growth protein pthr10333	protein	MGI	2810011M06Rik Ing1l P33ING2
Stra8	stimulated by retinoic acid gene 8	Mus musculus		protein	MGI	
Zbtb16	zinc finger and BTB domain containing 16	Mus musculus	zinc finger protein pthr11389	protein	MGI	Green's luxoid PLZF Zfp145
Etv5	ets variant 5	Mus musculus	ets pthr11849	protein	MGI	1110005E01Rik 8430401F14Rik erm

Functional analysis relates your hits to a set of pre-defined functional groups

A4galt

Atl1

Cdk19

Cdon

Cecr2

Etv5

Flywch1

Gnpda2

Hoxc4

Ing2

ligp1

Map3k9

Mypop

Rnf6

Serinc1

Stra8

Trp73

Zbtb16

Functional analysis relates your hits to a set of pre-defined functional groups

A4galt

Atl1

Cdk19

Cdon

Cecr2

Etv5

Flywch1

Gnpda2

Hoxc4

Ing2

ligp1

Map3k9

Mypop

Rnf6

Serinc1

Stra8

Trp73

Zbtb16

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.

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

There are many sources of functional gene lists

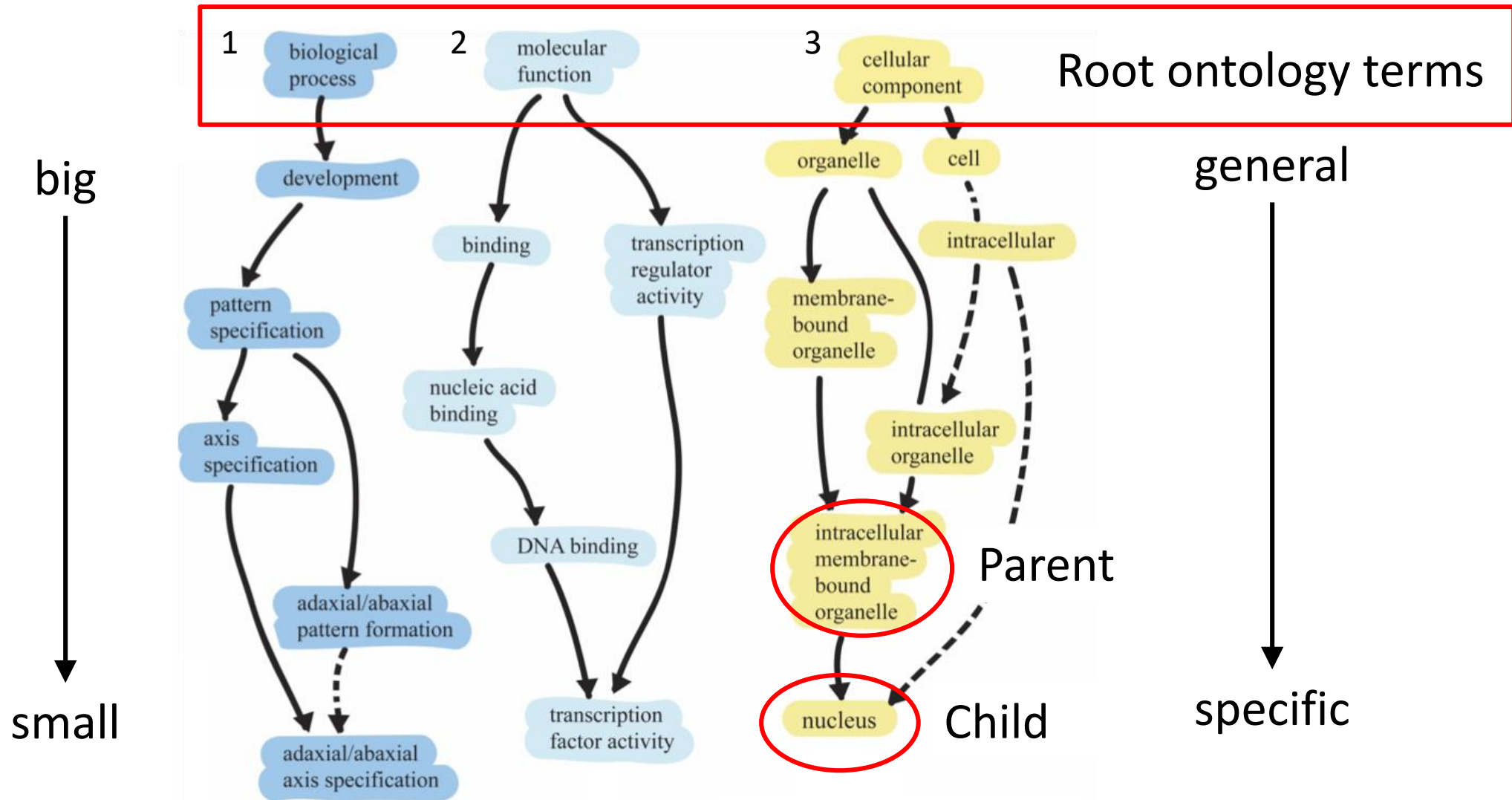
- Human curated
 - Gene Ontology
 - Biological Pathways
- Domains / Patterns
 - Protein functional domains
 - Transcription factor regulated
- Experimental
 - Co-expressed genes
 - Interactions
 - Hits from other studies

Gene Ontology is a human curated functional database



GENEONTOLOGY
Unifying Biology

GO has three domains and a hierarchical structure



Genes are placed into each domain as specifically as possible

Nanog homeobox [Source:HGNC Symbol;Acc:HGNC:20857]

- Cellular Component
 - GO:0005634 nucleus
 - GO:0005654 nucleoplasm
 - GO:0005730 nucleolus
- Molecular Function
 - GO:0003677 DNA binding
 - GO:0003700 transcription factor activity, sequence-specific DNA binding
 - 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 polymerase II promoter
 - GO:1903507 negative regulation of nucleic acid-templated transcription

Annotations come with evidence

- Experimental Evidence
 - Inferred from Experiment (EXP)
 - Inferred from Direct Assay (IDA)
 - Inferred from Physical Interaction (IPI)
 - Inferred from Mutant Phenotype (IMP)
 - Inferred from Genetic Interaction (IGI)
 - Inferred from Expression Pattern (IEP)

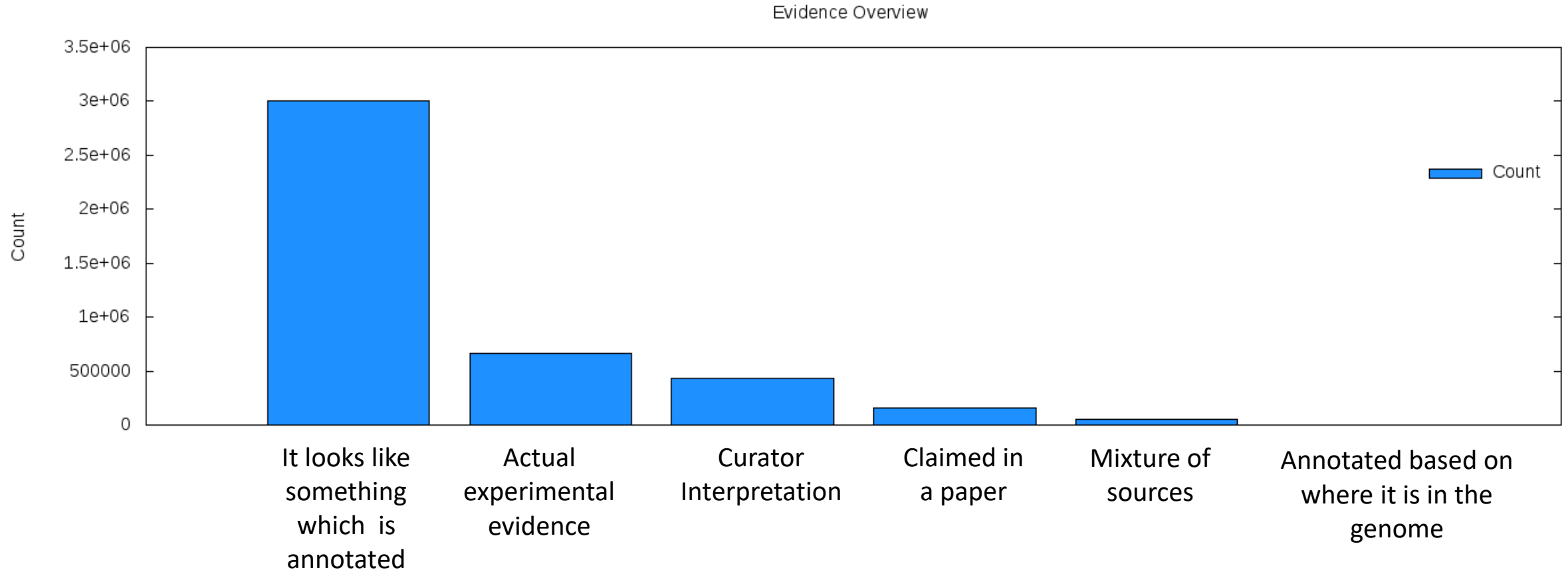
Annotations come with evidence

- Computational Evidence
 - Inferred from Sequence or structural Similarity (ISS)
 - Inferred from Sequence Orthology (ISO)
 - Inferred from Sequence Alignment (ISA)
 - Inferred from Sequence Model (ISM)
 - Inferred from Genomic Context (IGC)
 - Inferred from Biological aspect of Ancestor (IBA)
 - Inferred from Biological aspect of Descendant (IBD)
 - Inferred from Key Residues (IKR)
 - Inferred from Rapid Divergence (IRD)
 - Inferred from Reviewed Computational Analysis (RCA)

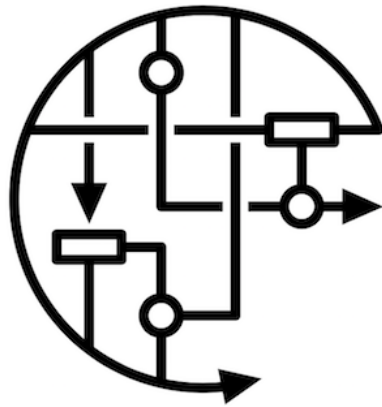
Annotations come with evidence

- Publications
 - Traceable Author Statement (TAS)
 - Non-traceable Author Statement (NAS)
- Curators
 - Inferred by Curator (IC)
 - No biological Data available (ND)
- Automated assignment
 - Inferred from Electronic Annotation (IEA)

Annotations come with evidence



Pathway databases trace metabolic pathways and their regulation



WIKIPATHWAYS
Pathways for the People



Protein Domain databases map out functional subdomains within proteins

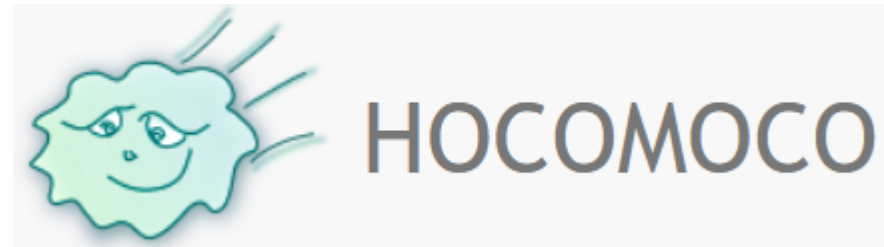
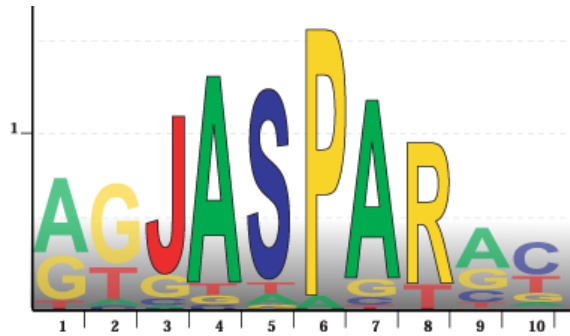
Pfam



InterPro
Protein sequence analysis & classification



Transcription Factor databases group genes by the motifs in their promoters

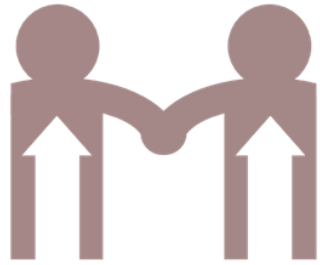


SwissRegulon

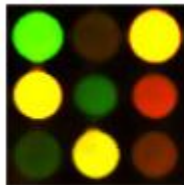
Swiss Institute of
Bioinformatics

AnimalTFDB3.0

Co-expression databases group genes which are expressed together

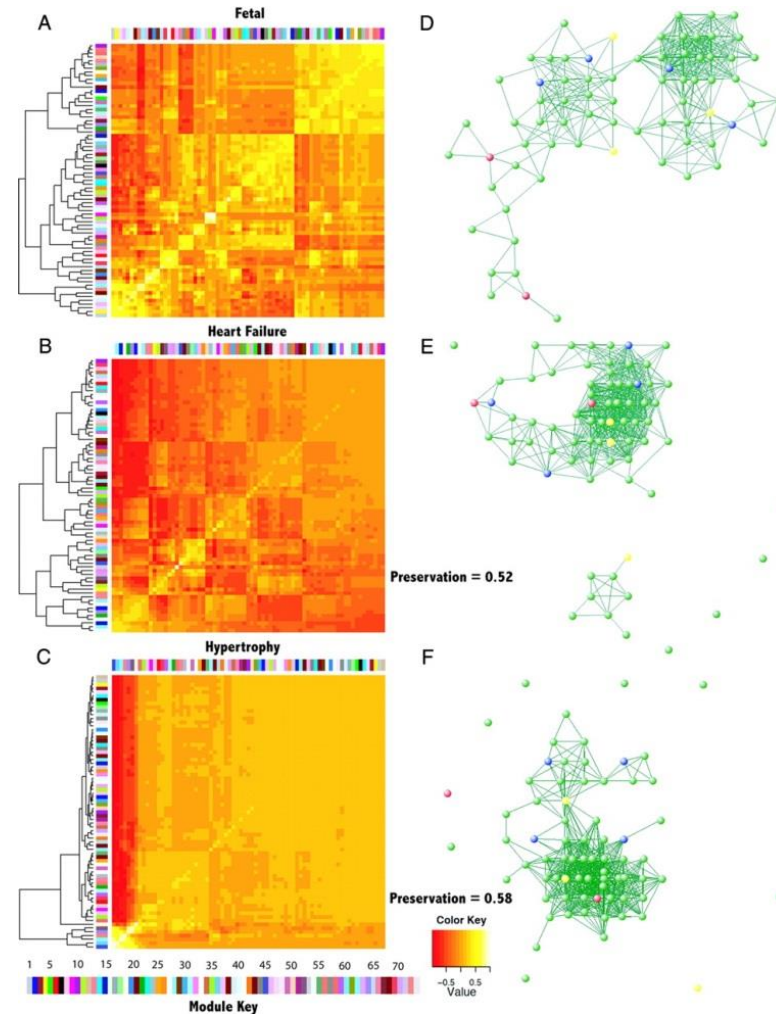


GeneFriends



Coexpedia

Powered by NETBIOLAB.org



Interaction databases map out physical interactions between genes and their products



Some databases collate gene sets from many different sources



Pathway Commons

Access and discover data integrated from public pathway and interactions databases.



MSigDB

Molecular Signatures
Database

Molecular Signatures Database v6.2

Testing for enriched gene sets

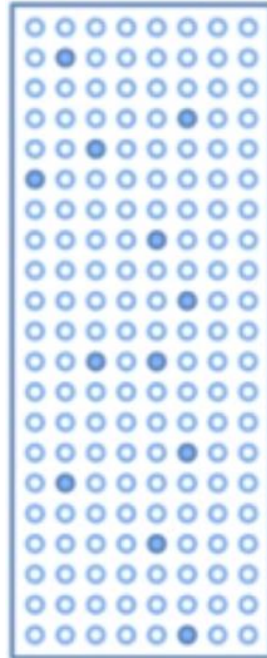
There are two basic ways to test for enrichment

- 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

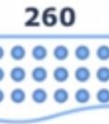
13,101 genes
on chip



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



Gene List



Related to
disease
 $260/747 =$
34.8%



Not related to
disease

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

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

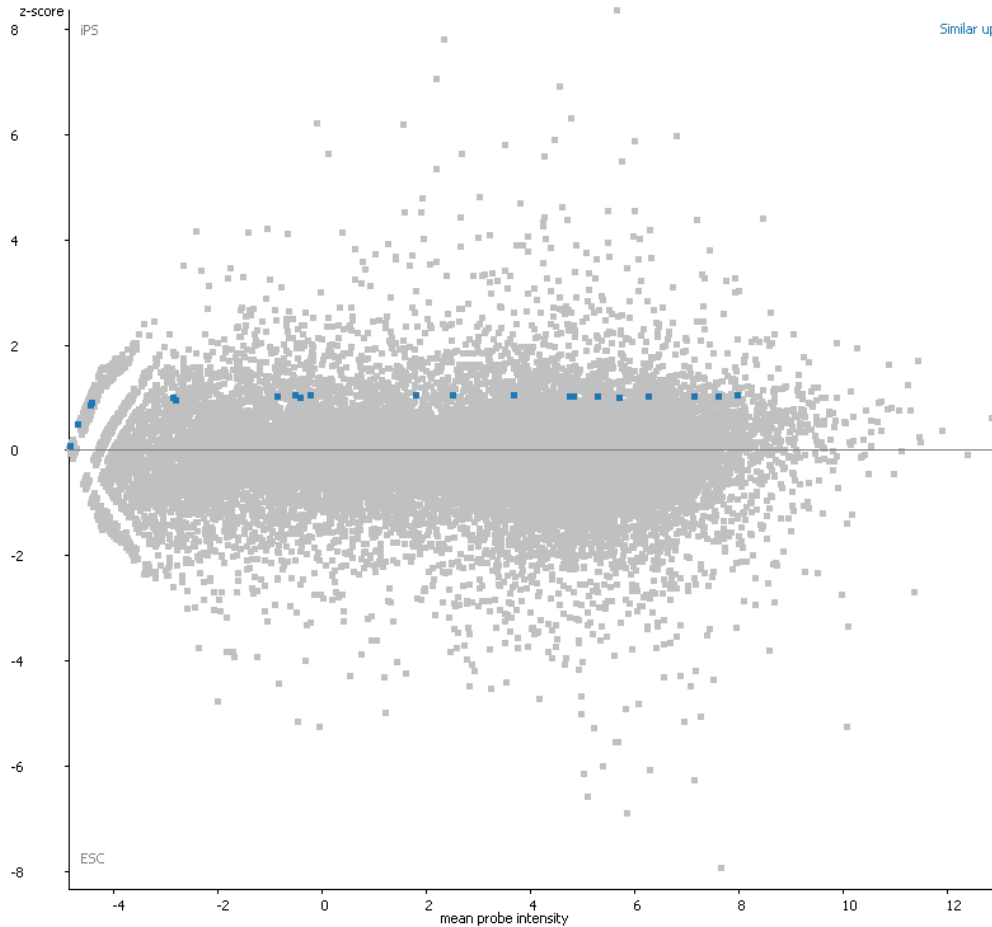
- Function X
 - Length=1 $p=0.60$
 - Length=2 $p=0.80$
 - Length=3 $p=0.30$
 - Length=4 $p=0.35$
 - Length=5 $p=0.40$
 - Length=6 $p=0.45$
 - Length=7 $p=0.05$
 - Length=8 $p=0.08$
 - Length=9 $p=0.10$

Quantitative Enrichment Analysis

Quantitative comparisons offer more power, if you have a suitable metric

- What quantitation can we use?
 - Differential p-value (normally $-10 \log(p)$)
 - Fold change
 - Absolute difference
- Measures often have odd distributions and biases
 - Z-scores
 - Ranks

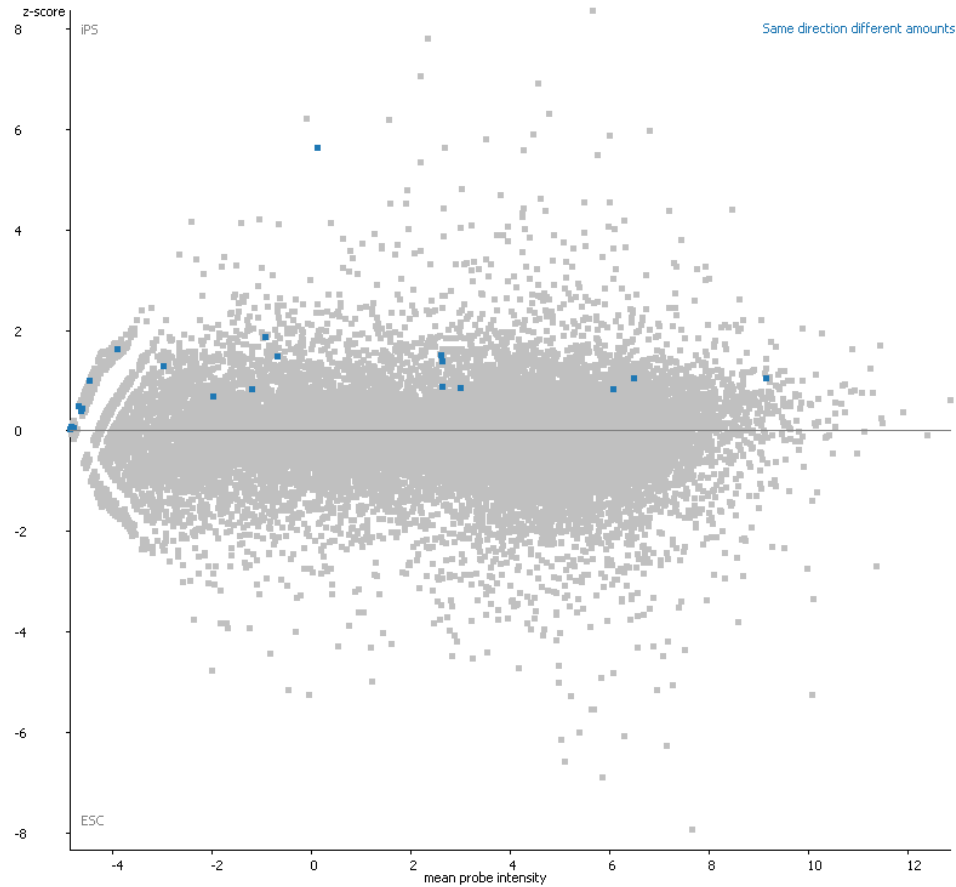
What kind of changes do we expect in an interesting category?



Student's T-test

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

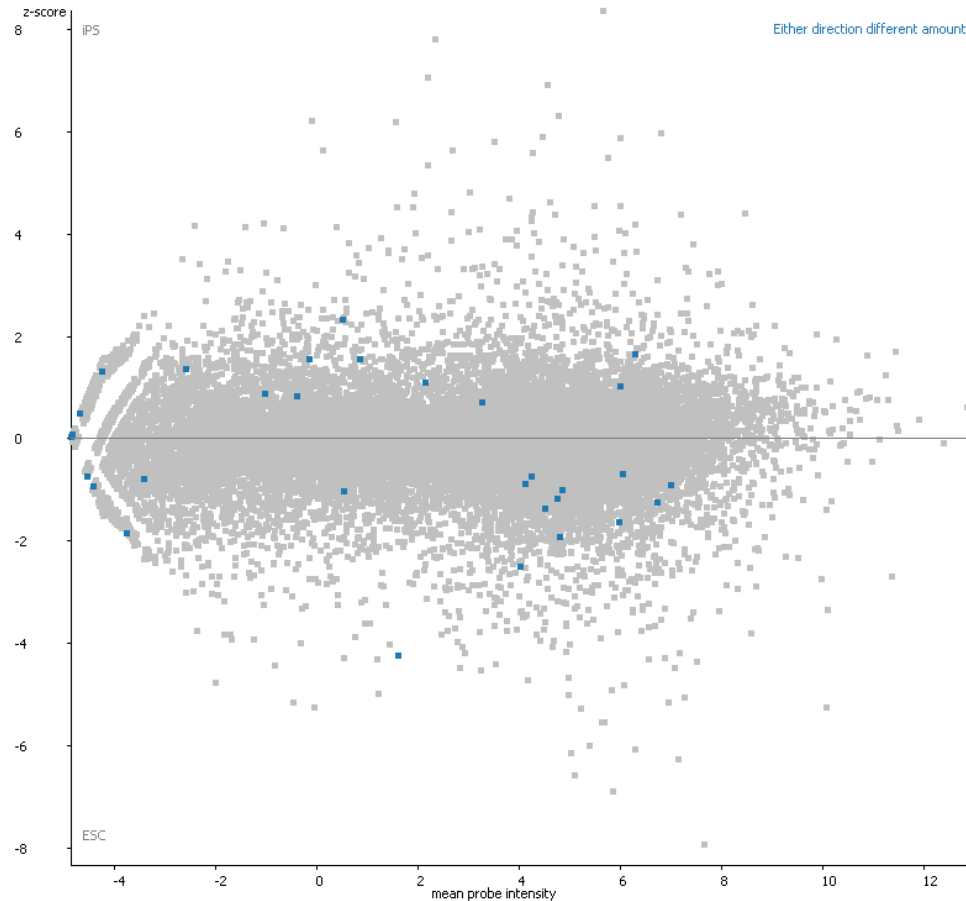
What kind of changes do we expect in an interesting category?



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?

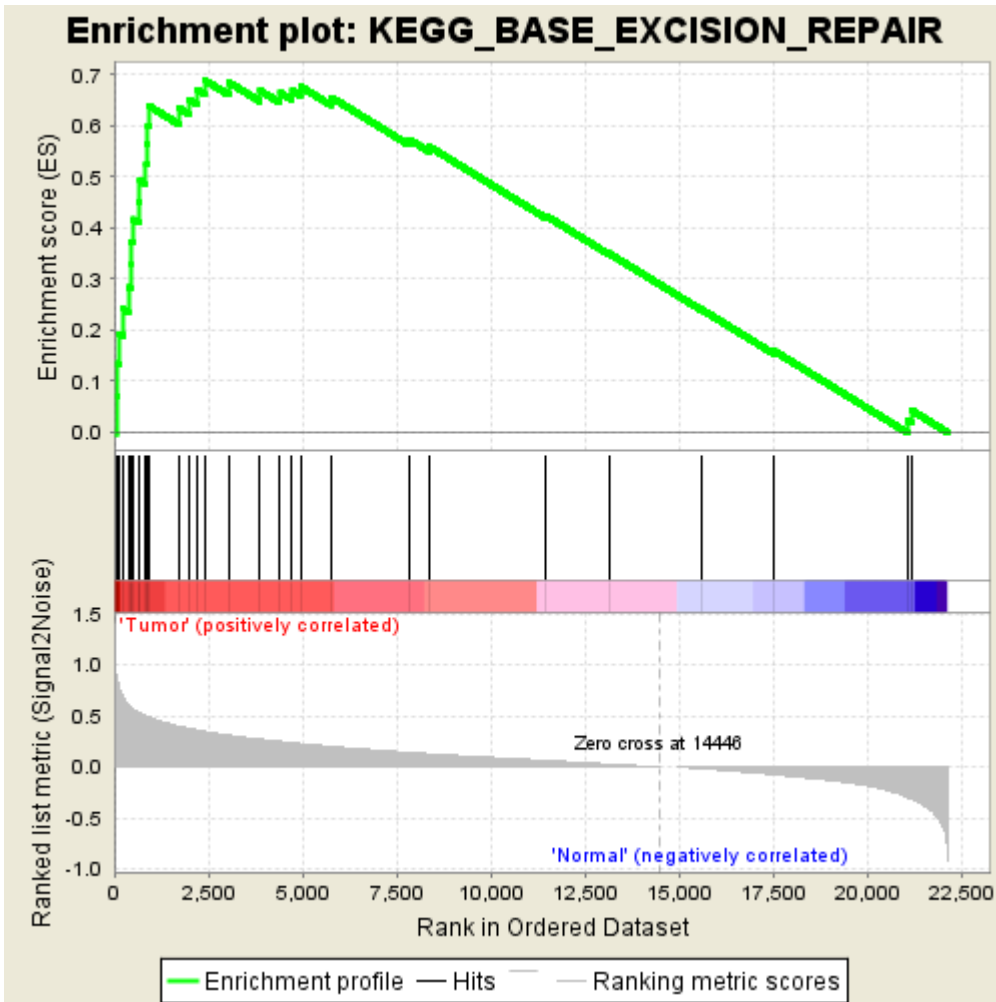
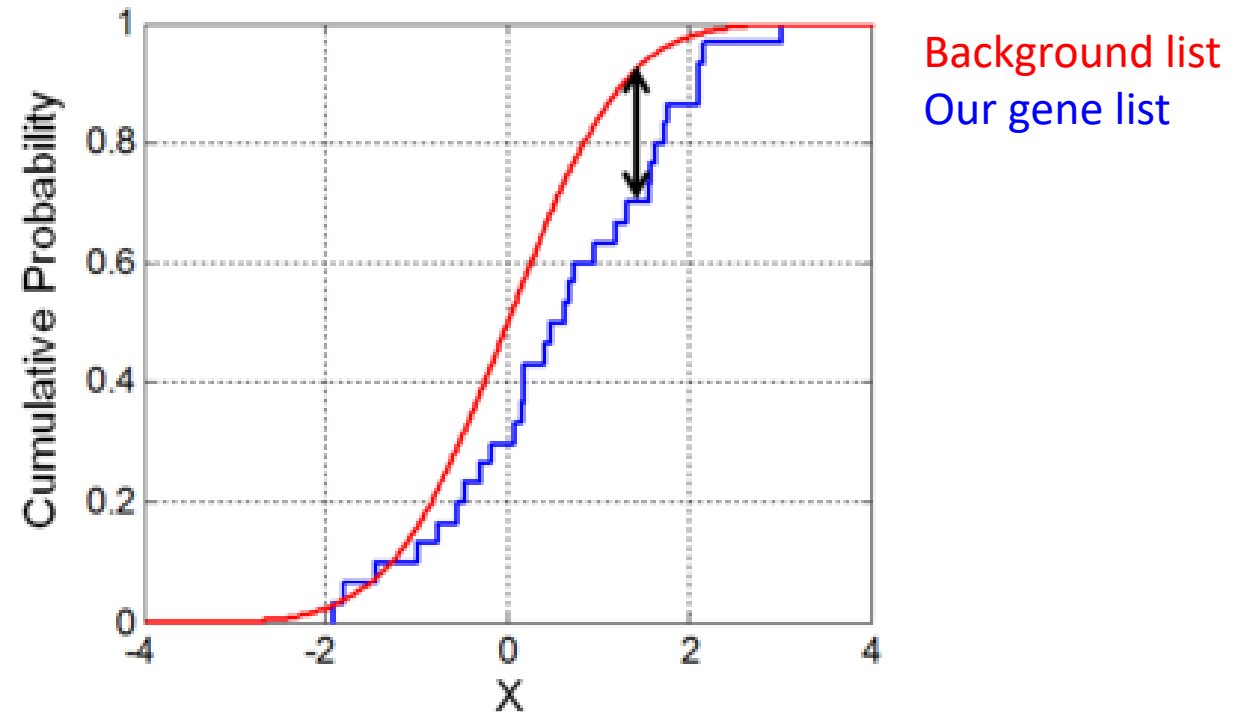


Absolute KS Test

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

Kolmogorov Smirnov

- Looks for the biggest point of difference between the background and test lists



Multiple testing correction

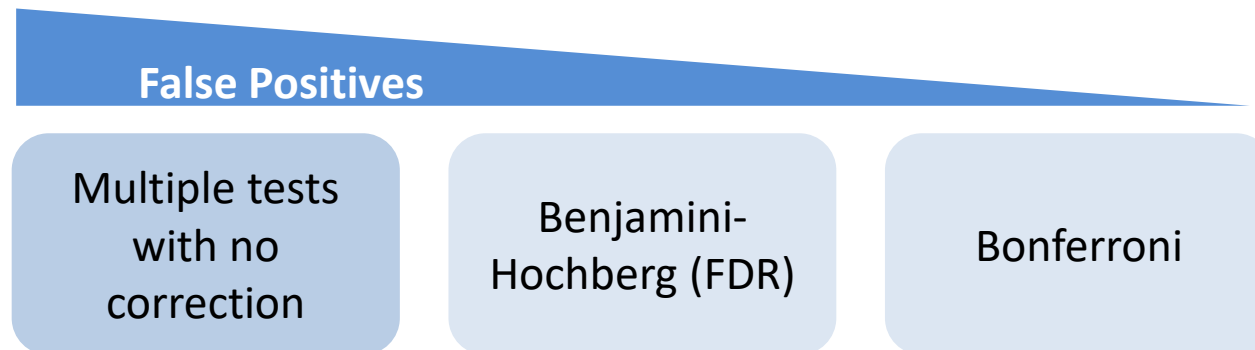
- More annotations/functions being tested = more chance of increase in false-positives

Bonferroni

- Significance level (e.g. 0.05) / number of tests = new threshold
- Over correction if tests are correlated

Benjamini-Hochberg

- Rank the p-values
- Apply more stringent correction to the most significant, and least stringent to the least significant p-values

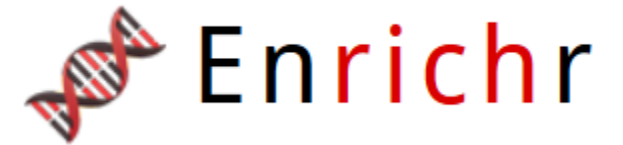


What do we get back from an enrichment test?

- 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:



GORILLA



Gene Ontology enRIchment anaLysis and visuaLizAtion tool



WebGestalt

g:GOS

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



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



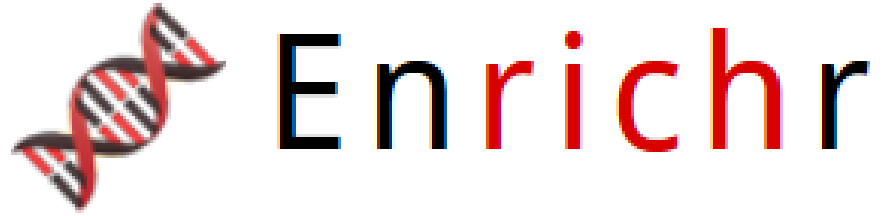
- Categorical or quantitative statistics
- Pathway focussed
- Simple submission interface (no custom background)
- Really nice visualisations

Goliath

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



- 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
 - No background list option
- Simple interactive visualisation
- Novel scoring scheme to rank hits



GORILLA

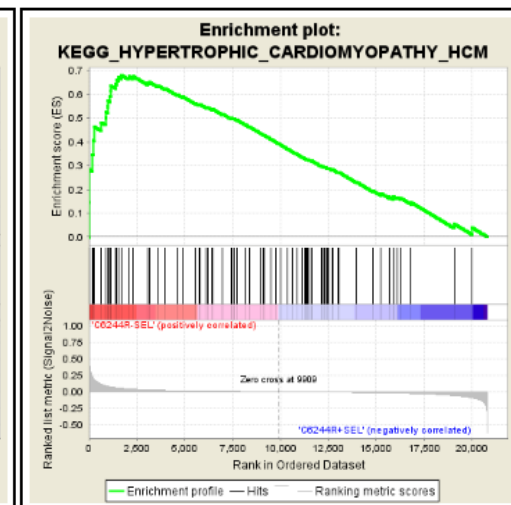
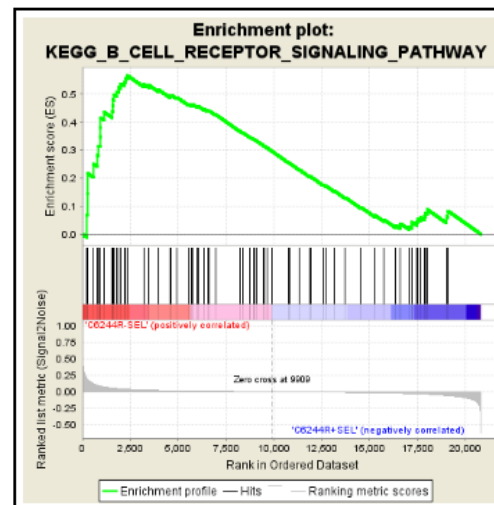
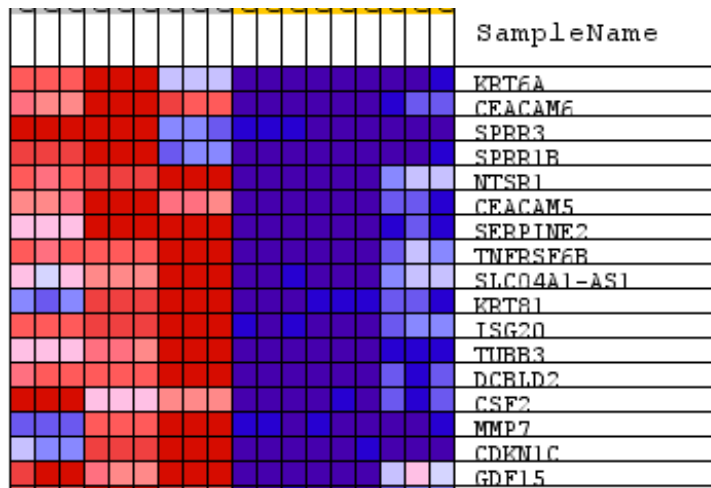


Gene Ontology enRIchment anaLysis and visuaLizAtion tool

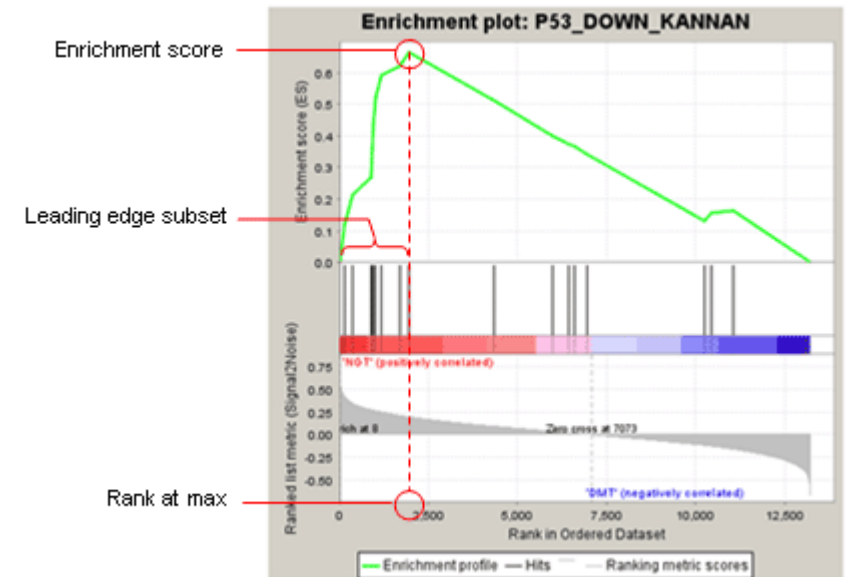
- Categorical or ranked analysis
- Mostly GO gene list support
- Interesting visualisation options

GSEA

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



- Genes ranked based on correlation to annotation groups
- Genes from a gene set placed onto the ranked lists
- Look for sets where there is unusual grouping at the top or the bottom of the list





SeqMonk Mapped Sequence Data Analyser

Version: 1.40.0

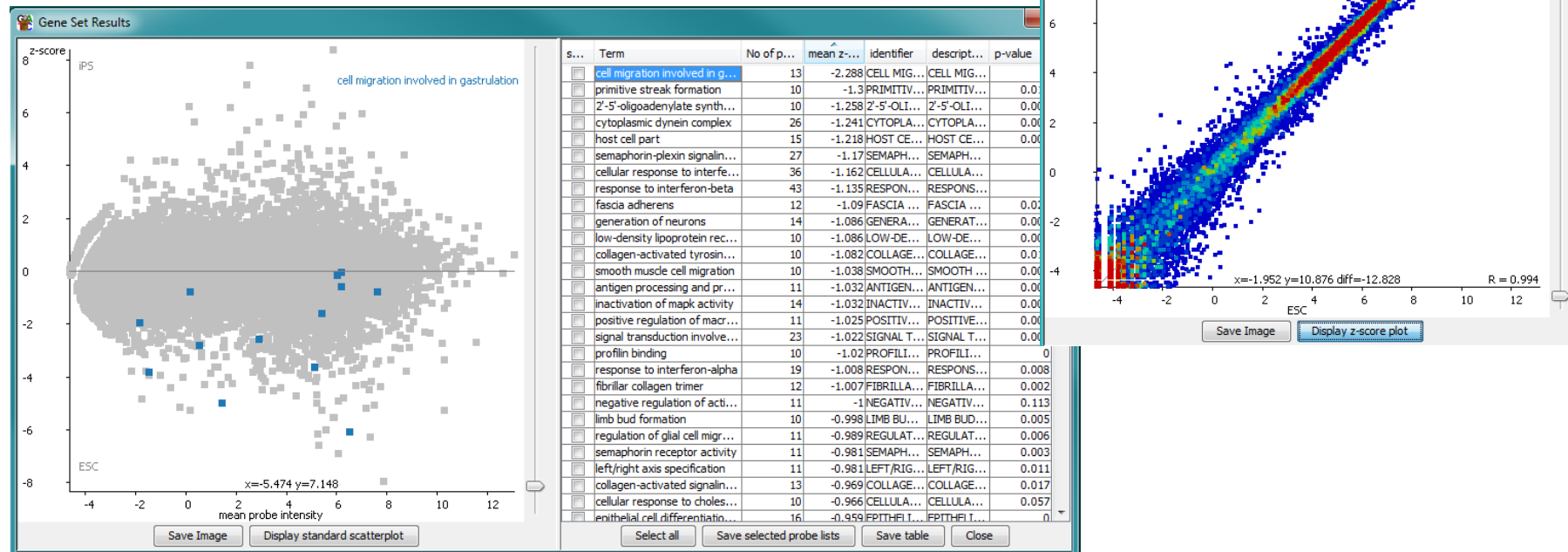
www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews, Laura Biggins Babraham Bioinformatics, 2006-17

Picard BAM/SAM reader ©The Broad Institute, 2009



- Quantitative enrichment of sequencing datasets
- Local Java application



Gene List Practical