Extracting Biological Information from Gene Lists

Simon Andrews, Laura Biggins, Boo Virk

simon.andrews@babraham.ac.uk
laura.biggins@babraham.ac.uk

v2024-11

Babraham Bioinformatics

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

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 <or= 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 Ox40l 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].

Harmonizome

GeneCards® THE HUMAN GENE DATABASE

wikigenes collaborative publishing

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.

ChatGPT

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.

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

My Hits

A4galt
Atl1
Cdk19
Cdon
Cecr2
Etv5

Flywch1
Gnpda2
Hoxc4
Ing2
Iigp1
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

reactome

InterPro
Protein sequence analysis & classification

STRING

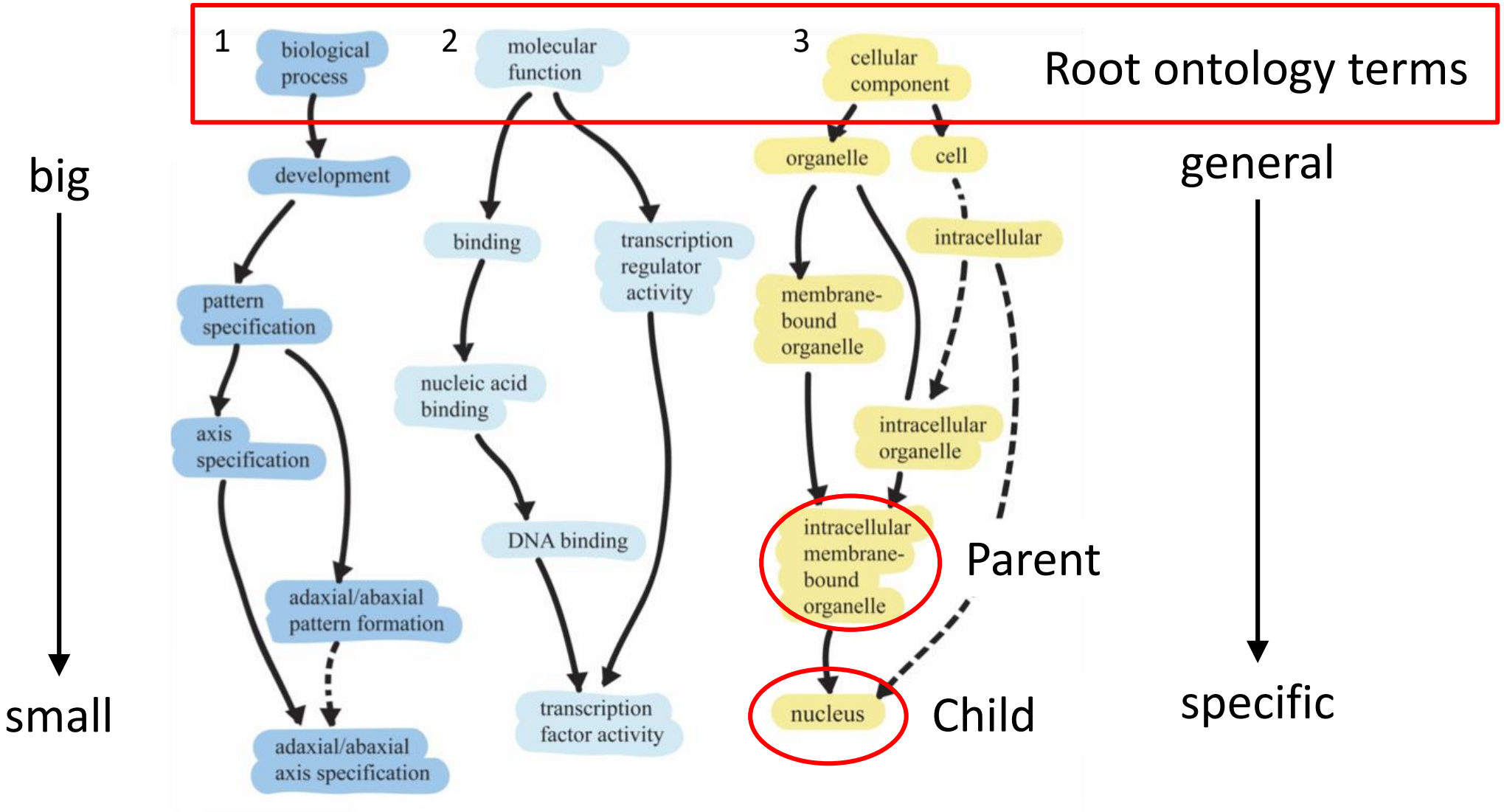
COXPRESdb

JASPAR

Gene Ontology is a human curated functional database

GENEONTOLOGY
Unifying Biology

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

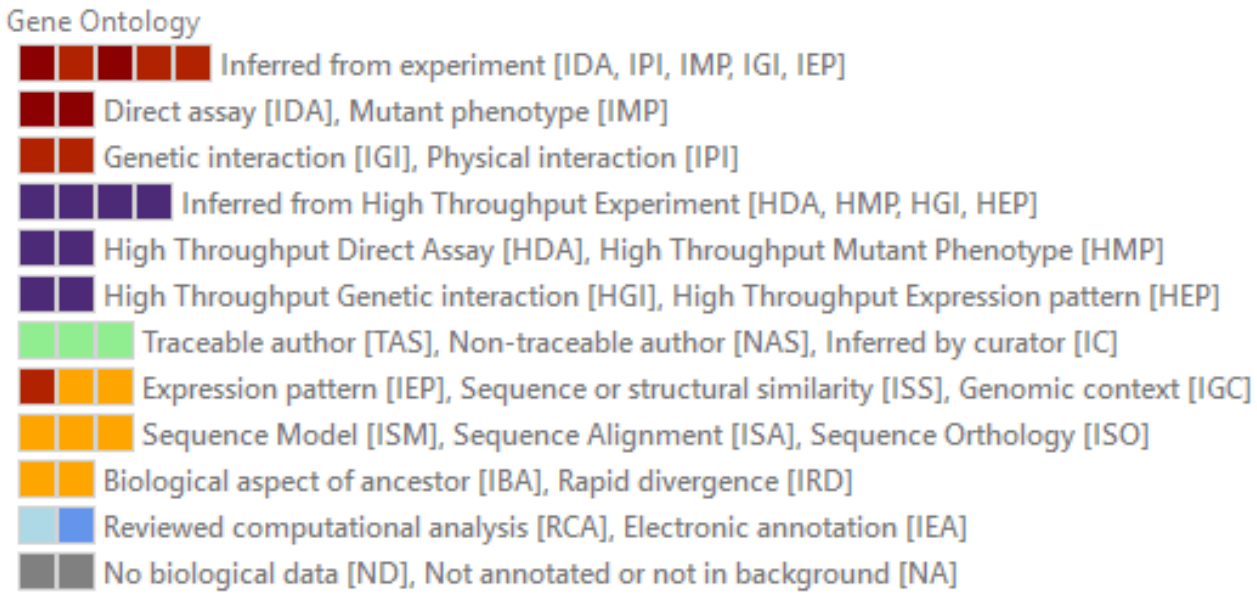
  GO:0045944 positive regulation of transcription from RNA pol2

  GO:1903507 negative regulation of nucleic acid-templated transcription

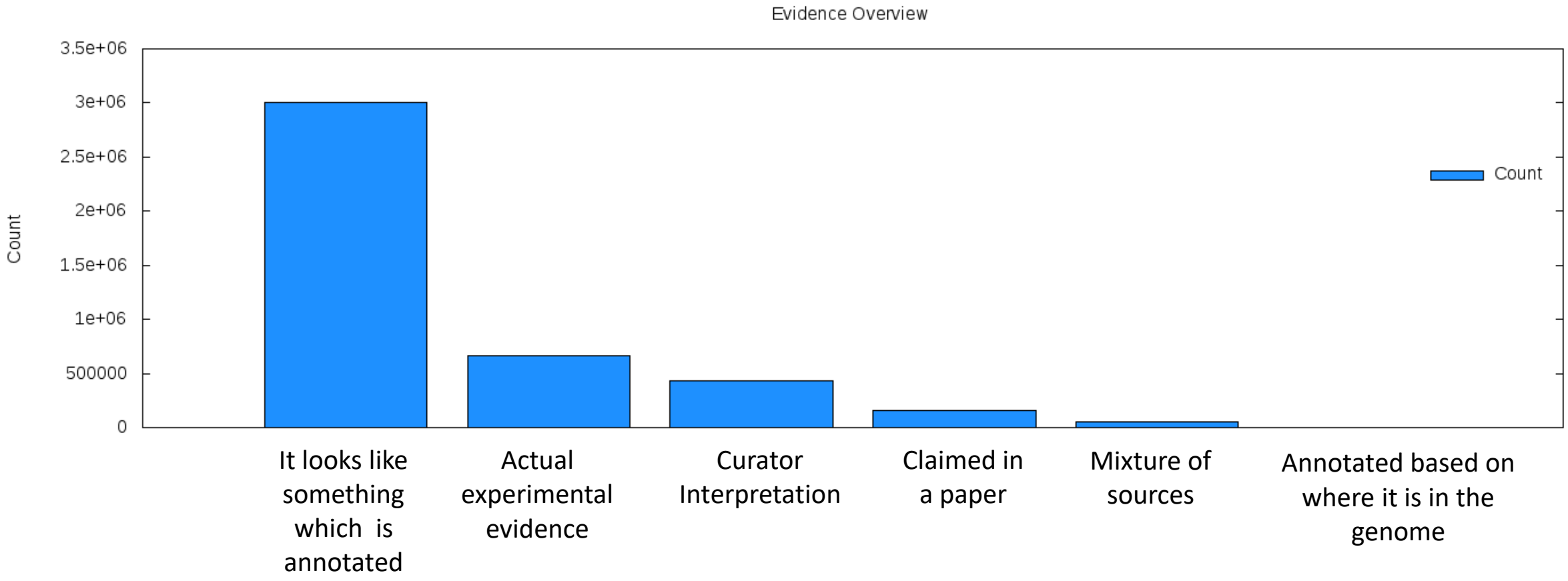
GO Annotations come with evidence

- 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)

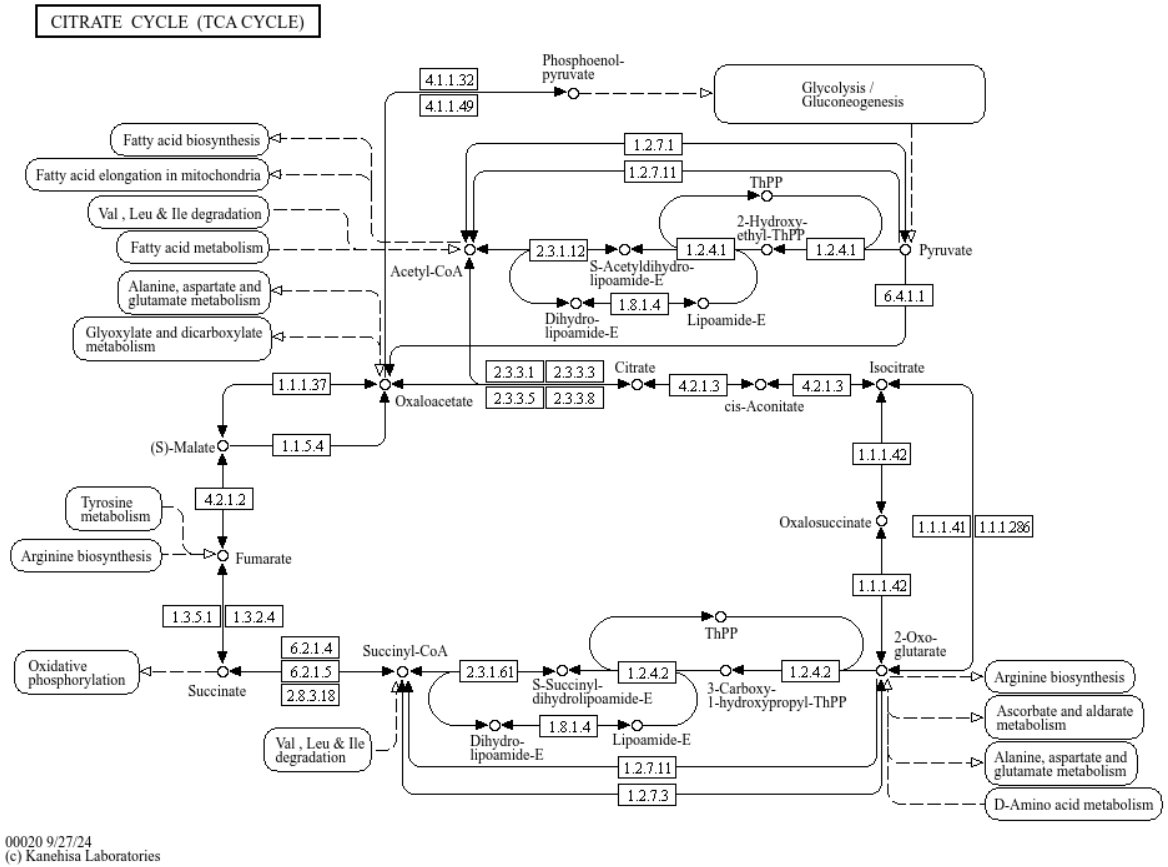
- Publications
- Curators



Annotations come with evidence



Pathway databases trace metabolic pathways and their regulation



Protein Domain databases annotate functional subdomains within proteins

PH domain

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

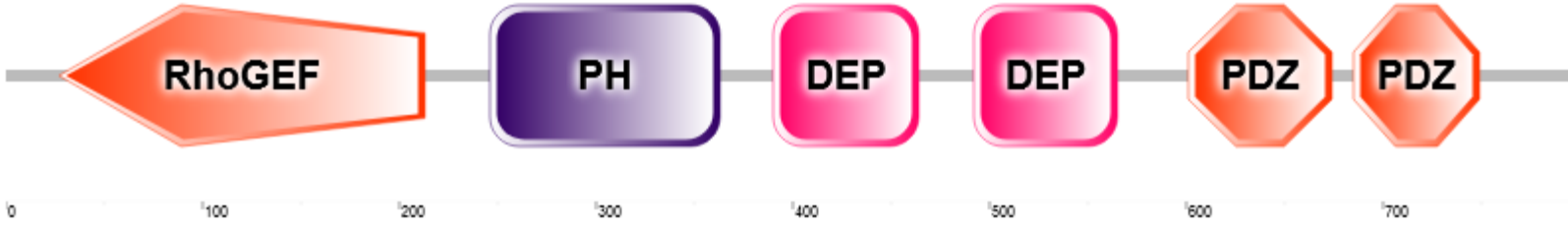
Position:	**246** to **363**
E-value:	**9.70401799592535e-12** (HMMER2)

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

+ = – Introns SAVE



Transcription Factor databases group genes by the motifs in their promoters

Swiss Regulon

JASPAR

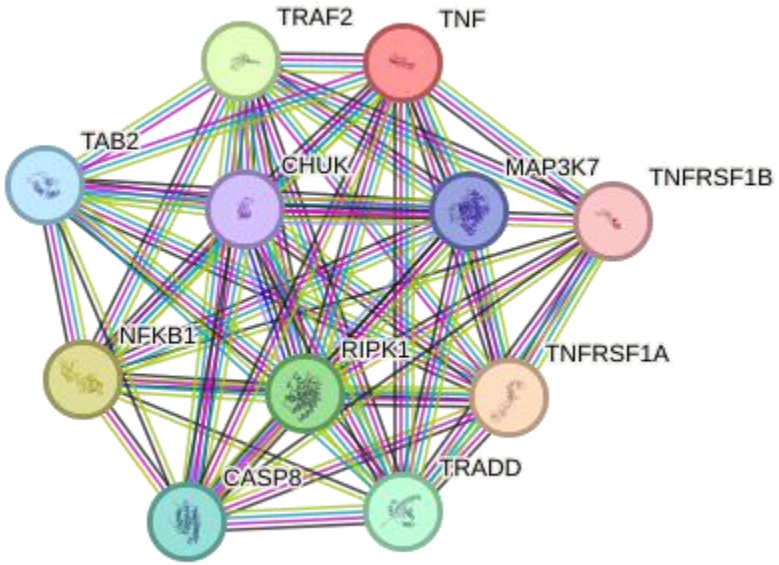
Factorbook

AnimalTFDB v4.0

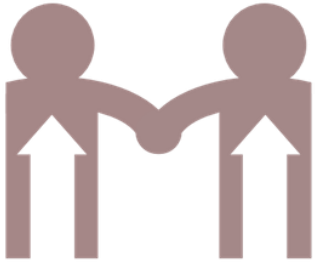
HOCOMOCO

Interaction databases map out interactions between genes / proteins

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



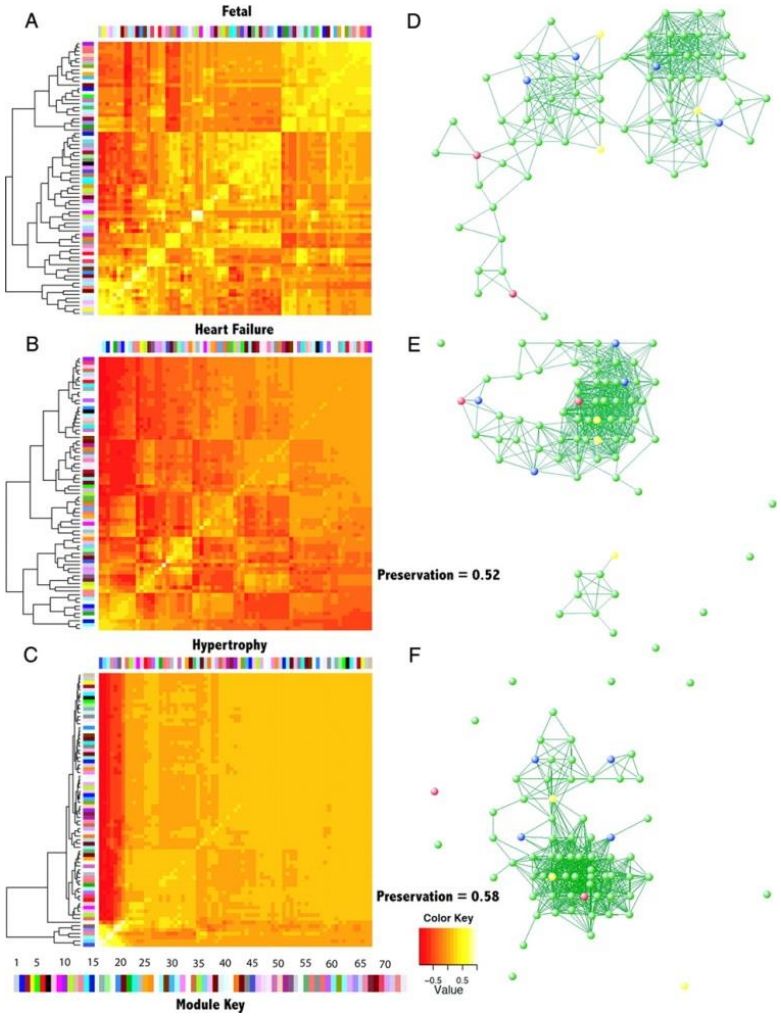
Co-expression databases group genes which are expressed together



GeneFriends

Coexpedia

Powered by NETBIOLAB.org



MSigDB

Molecular Signatures Database

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.

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.

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

Testing for enriched gene sets

Babraham Bioinformatics

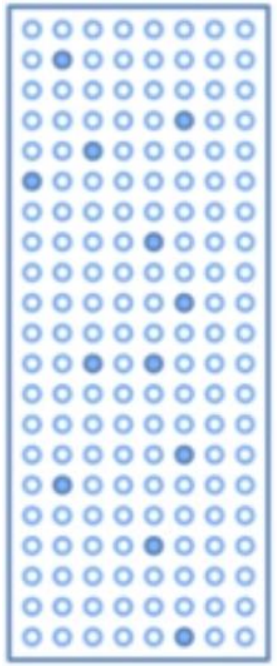
There are two basic ways to test for enrichment

- 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

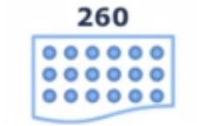
Categorical tests for enrichment

13,101 genes on chip

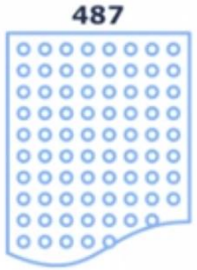


3005 genes related to disease 3005/13,101= 23.1%

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
```

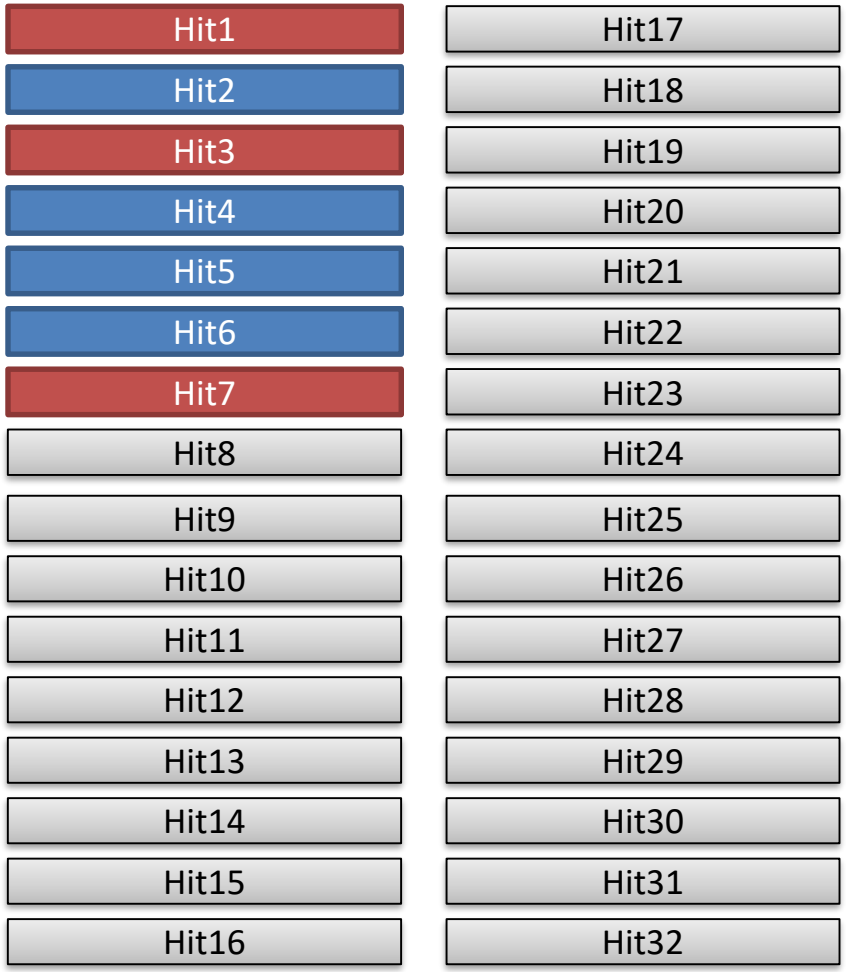
(260/487) / (3005/10096)

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

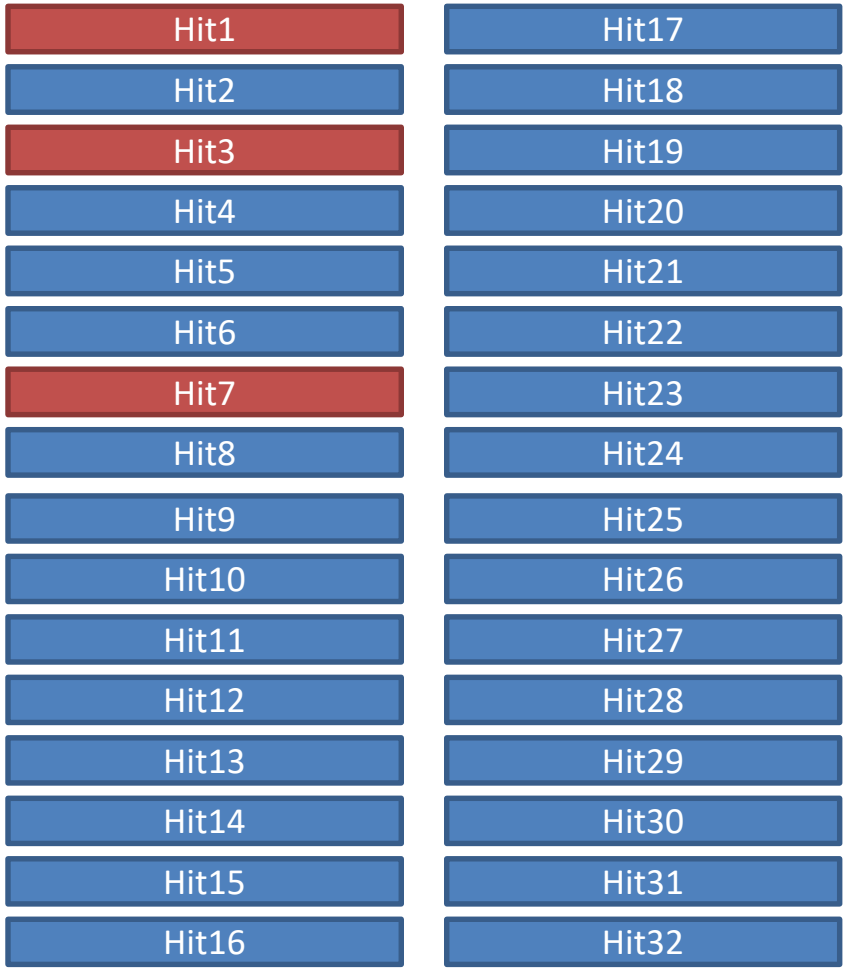
- Function X
  - 3 hits out of 32 in 'interesting' list
  - Not significant (p=0.07)

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



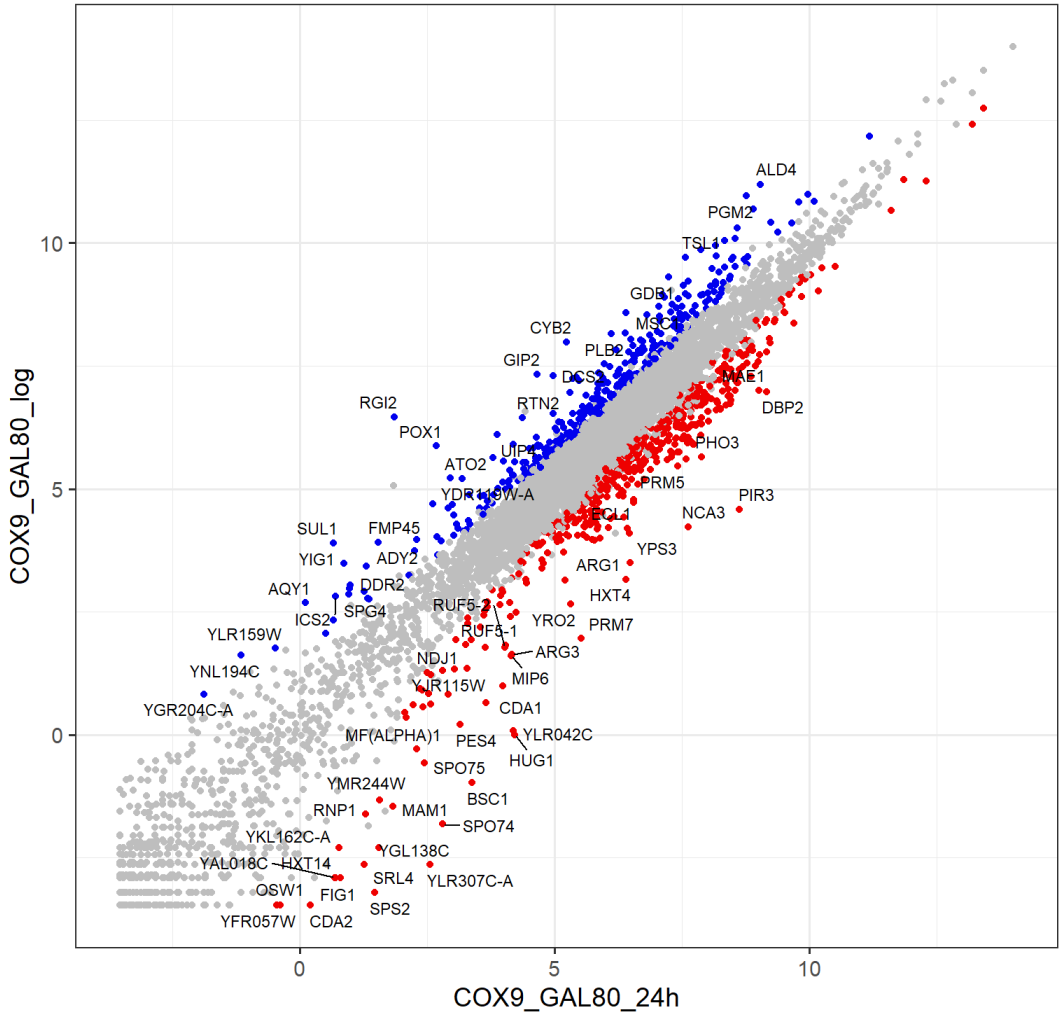
- Function X
  - 3 hits out of 7 in 'interesting' list
  - Significant (p=0.02)

Ordered, but not quantitative lists allow sequential categorical analysis



- 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



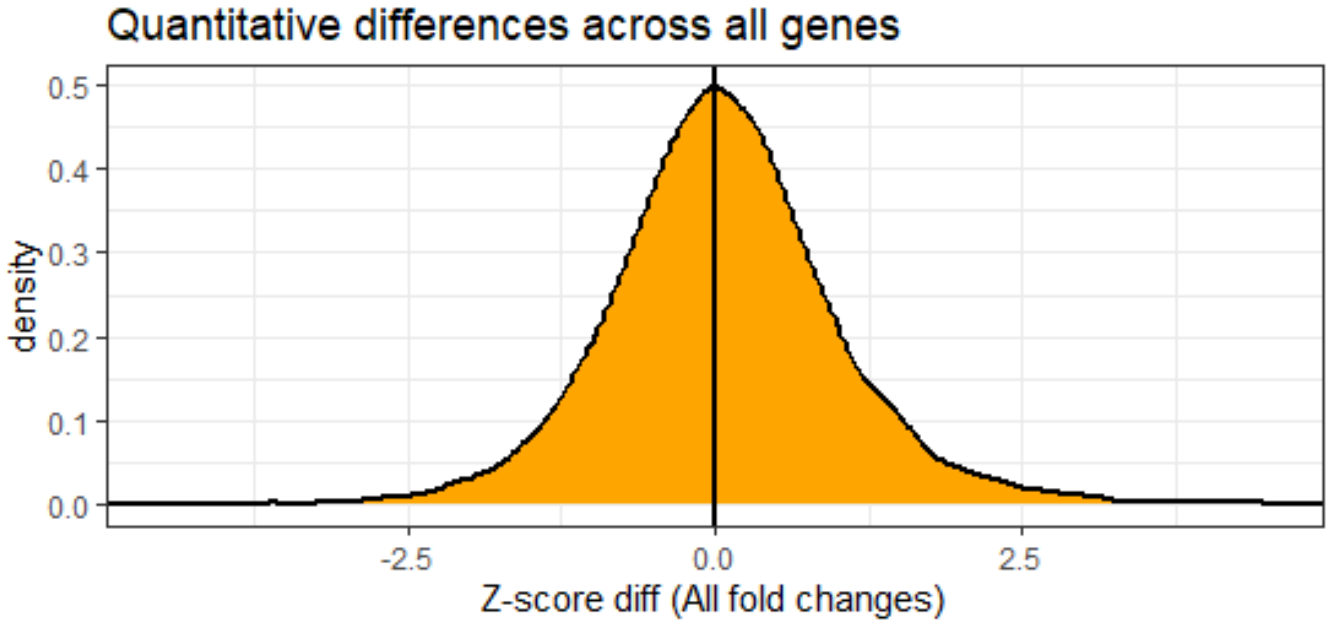
- One search or two?
  - One search
    - Higher power (more genes)
    - Lower enrichment
    - Mixed effects (pathways)
  - Two searches
    - Easier interpretation
    - Less power
    - Higher enrichment

Quantitative Enrichment Analysis

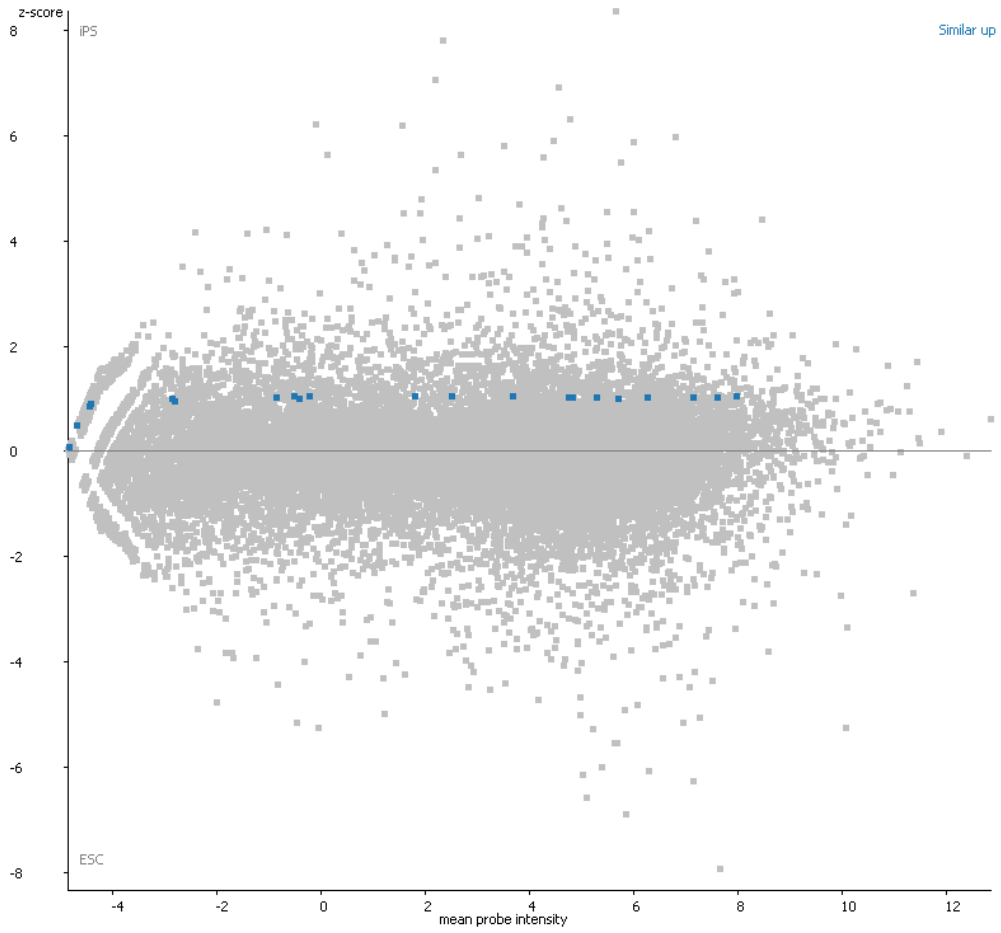
Babraham Bioinformatics

Quantitative comparisons can offer more power

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



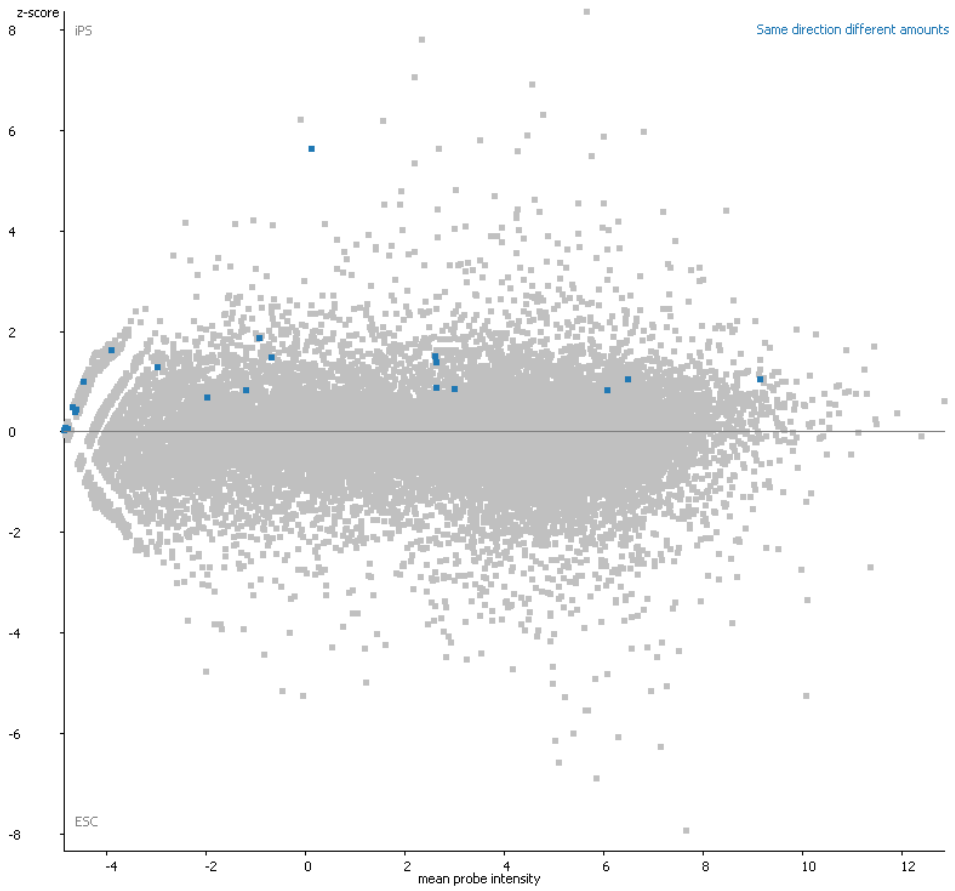
What kind of changes do we expect in an interesting category?



Student's T-test

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

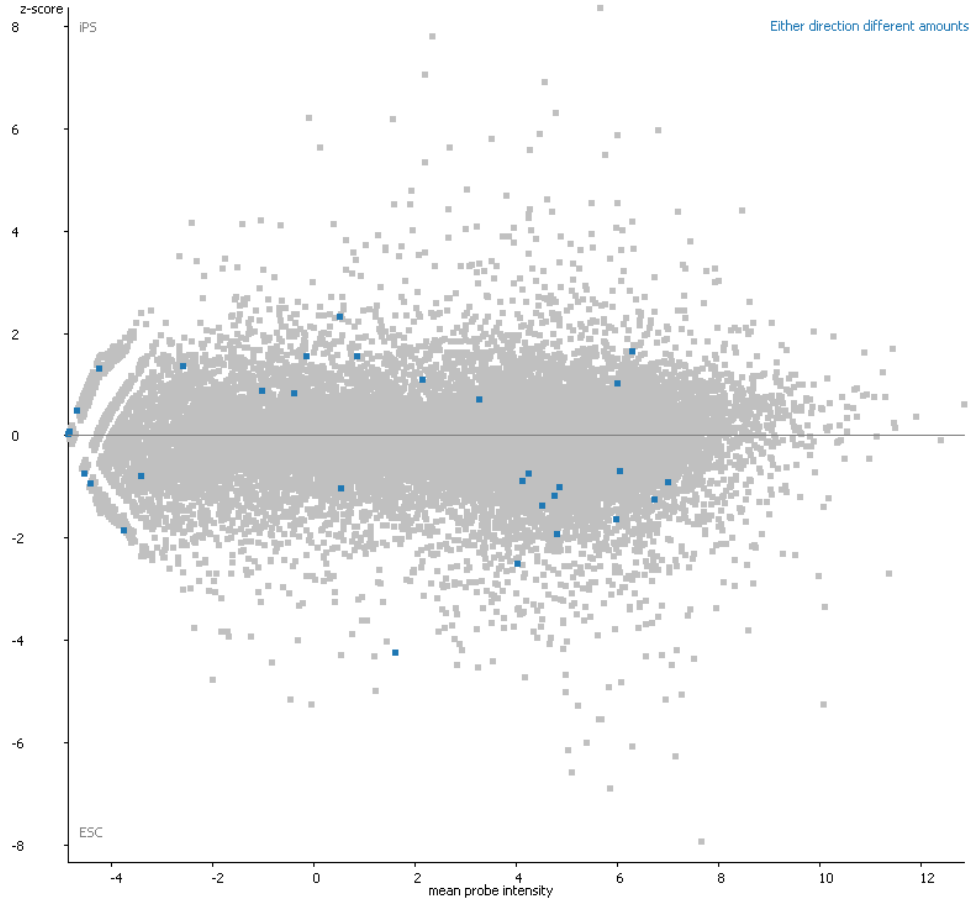
What kind of changes do we expect in an interesting category?



Kolmogorov Smirnov Test

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

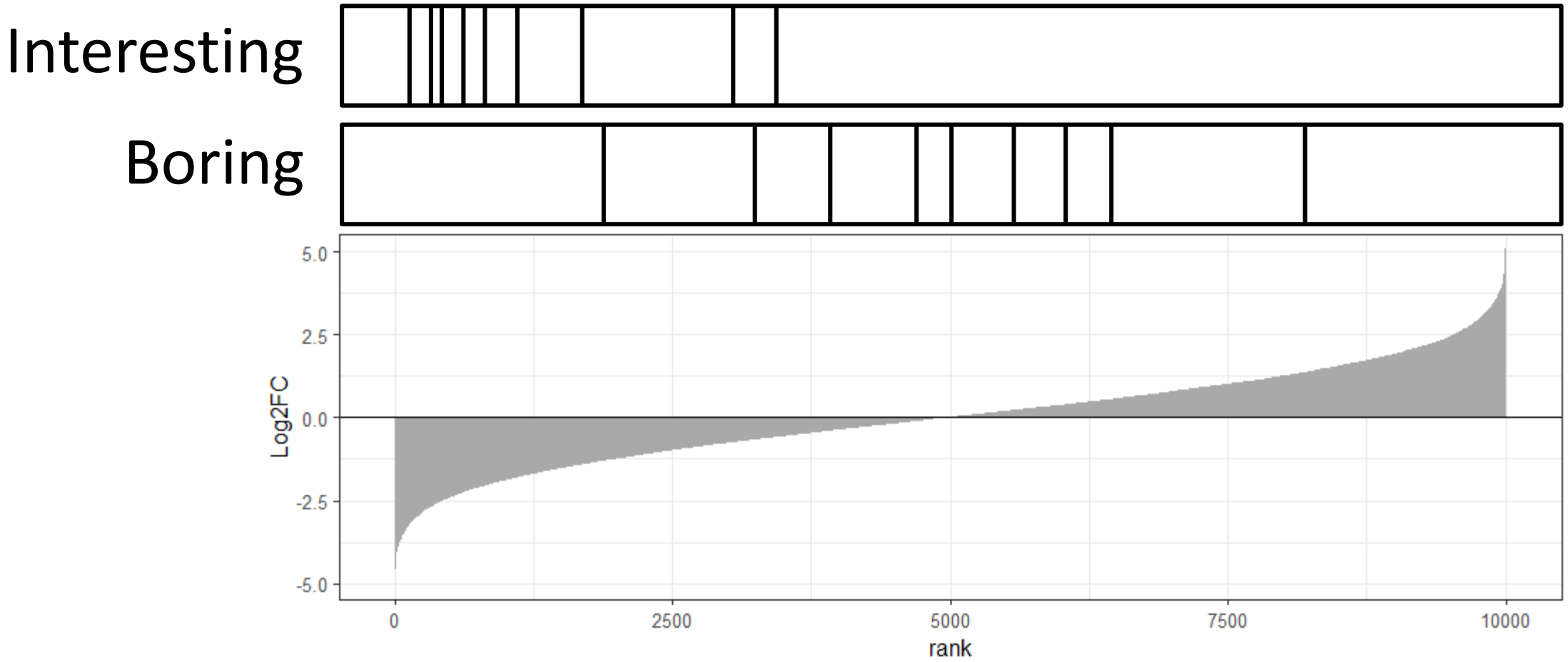
What kind of changes do we expect in an interesting category?



Absolute KS Test

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

GSEA statistics

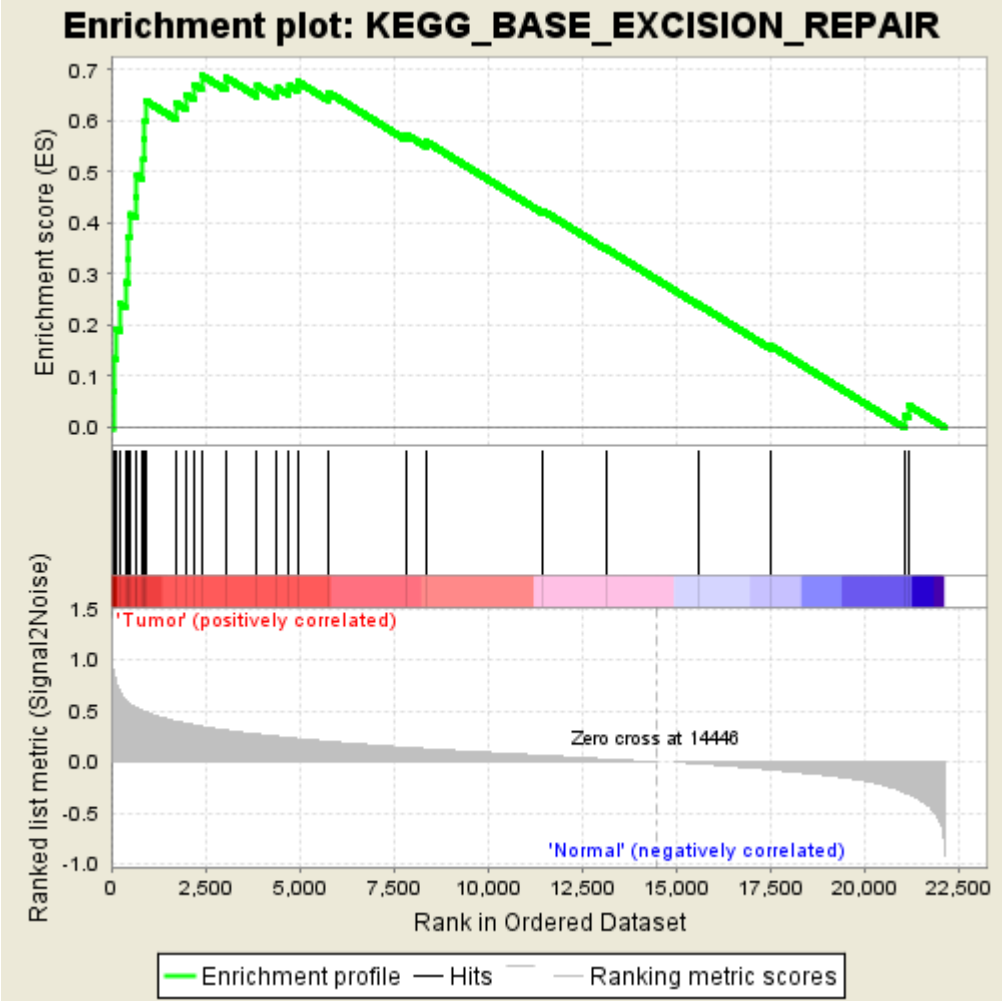


GSEA Statistics

- Keep a running total
- Start at the highest values
- If gene is in the set add value
- Otherwise subtract value

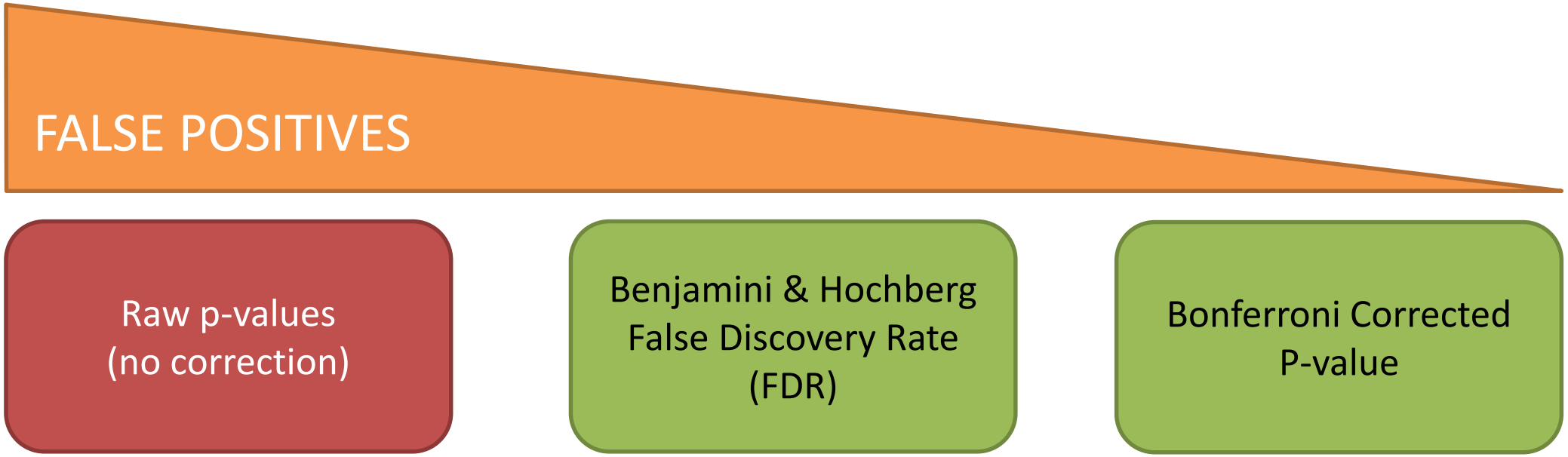
- Enrichment score is max score

- 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:

ToppGene

DAVID BIOINFORMATICS DATABASE

ShinyGO

Enrichr

PANTHER Classification System

g:Profiler

reactome

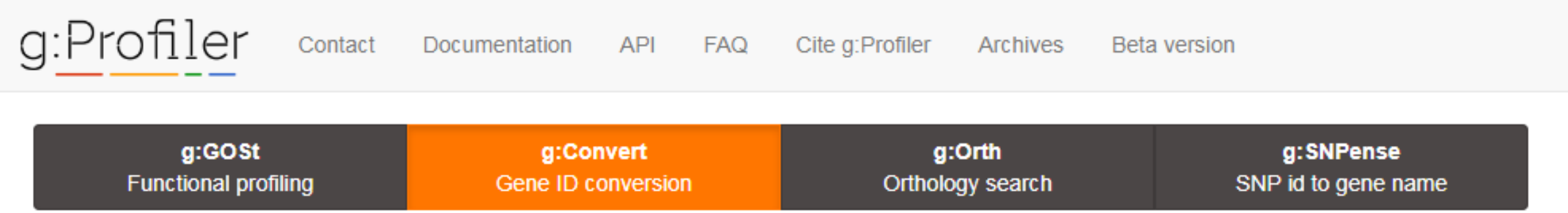
GOliath

GORILLA

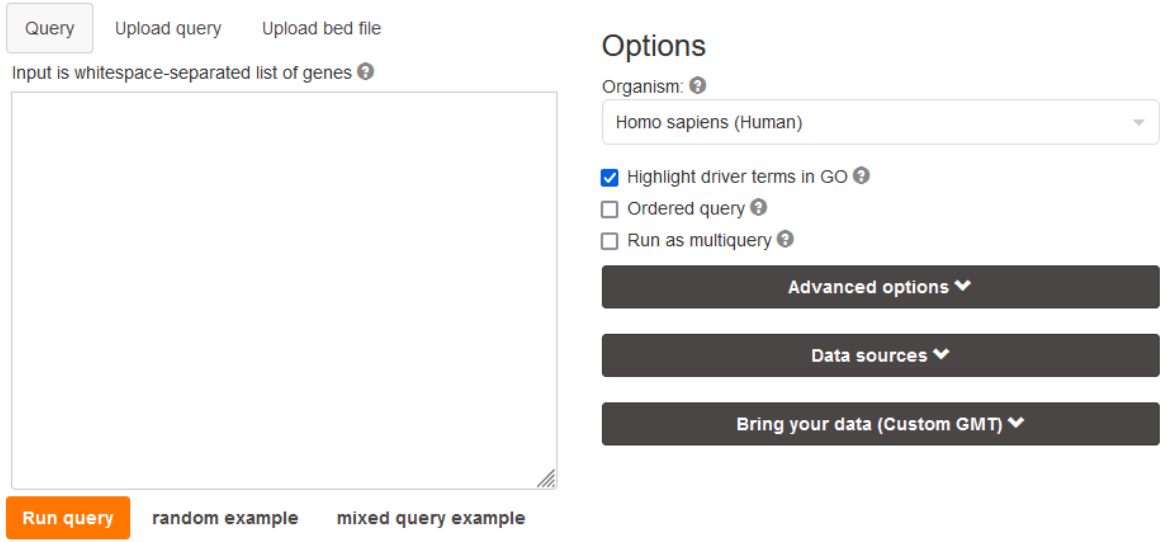
Gene Ontology enRIchment anaLysis and visuaLizAtion tool

GSEA Gene Set Enrichment Analysis

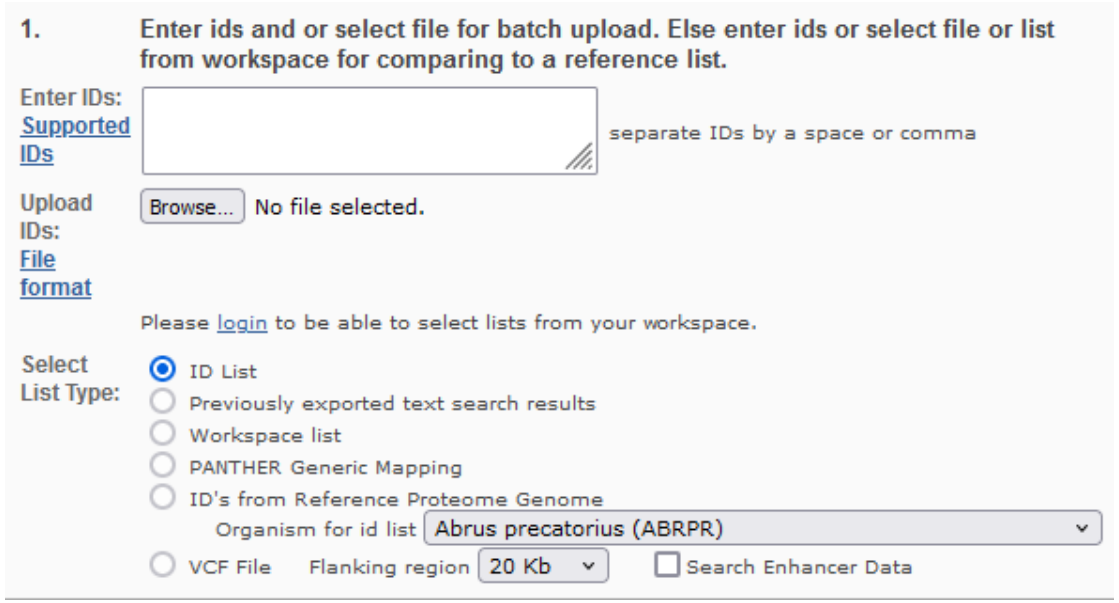
WebGestalt



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

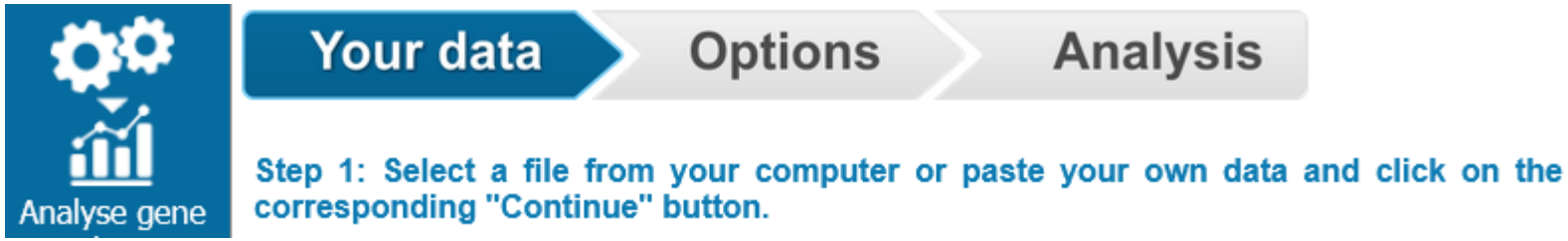


PANTHER
Classification System



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

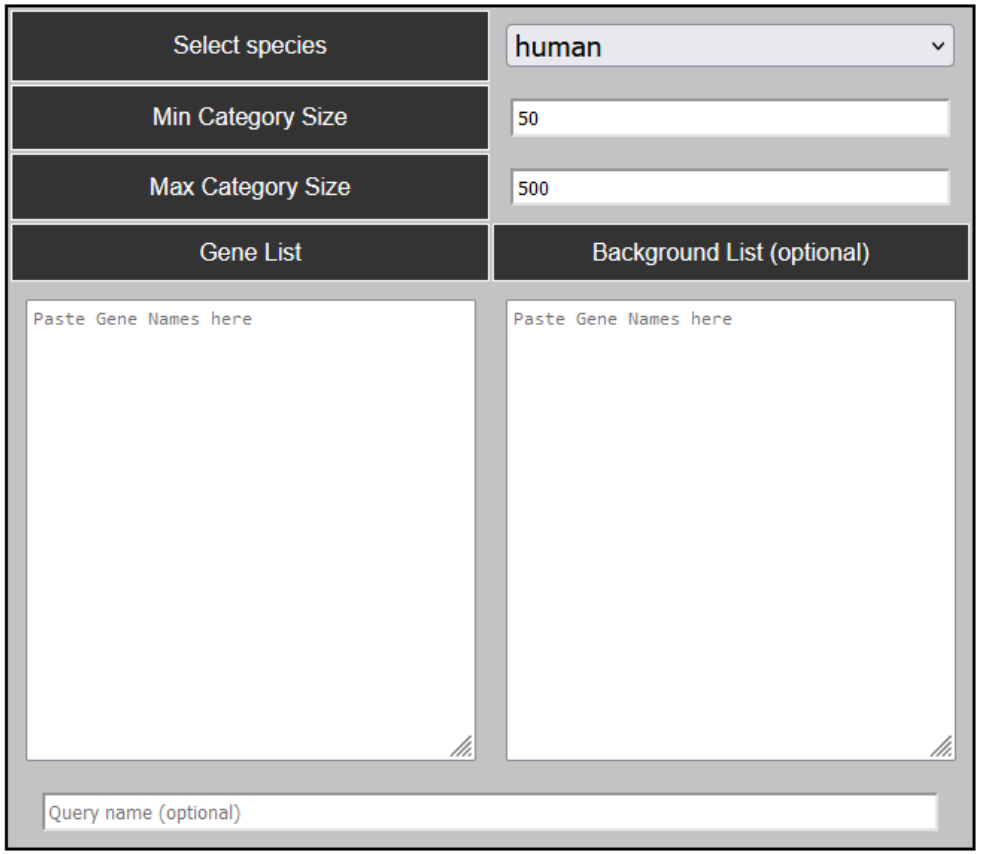
reactome



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

GOliath

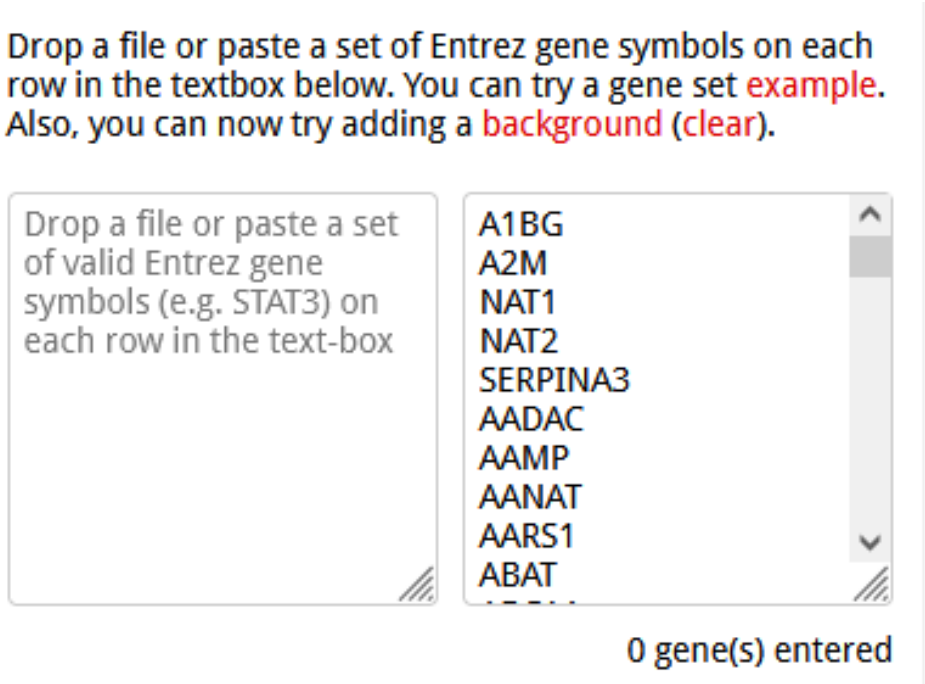
- Categorical statistics
- Limited species support
- Allows custom backgrounds
- Uses PathwayCommons gene sets
- Innovative detection and presentation of artefacts



- 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

Enrichr

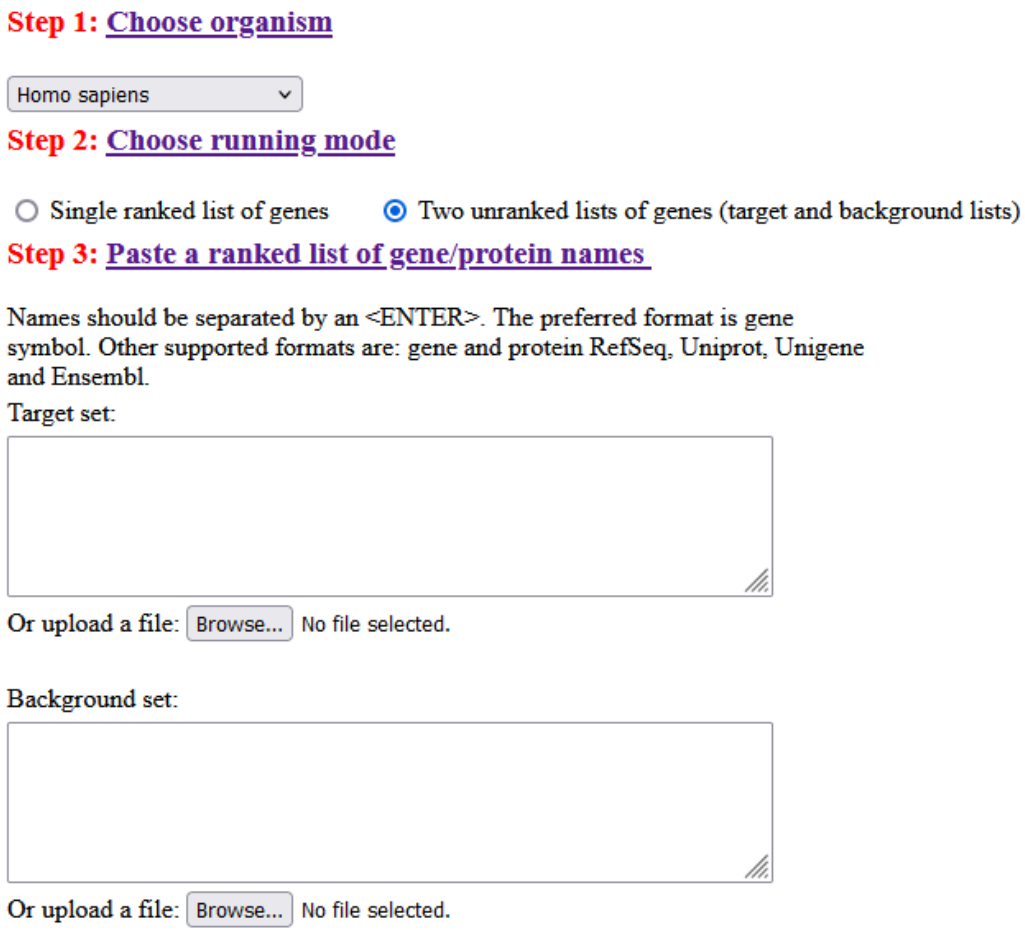
- Categorical Statistics
- Biggest selection of gene sets
- Simple interface, but limited options
  - No species information
- Simple interactive visualisation
- Novel scoring scheme to rank hits



GORILLA

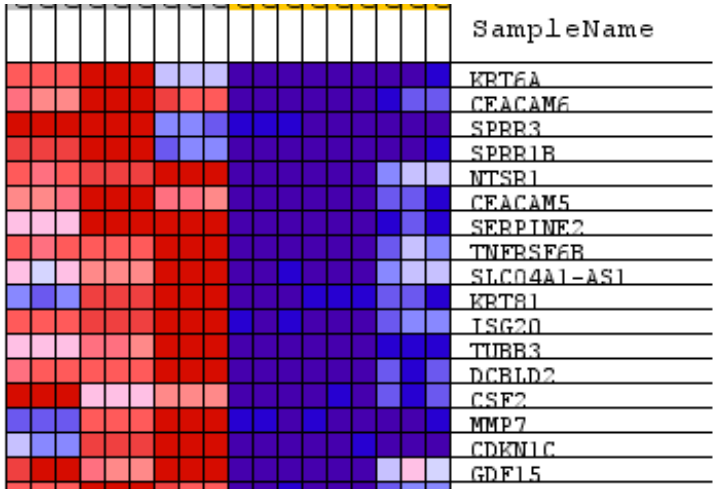
Gene Ontology enRIchment anaLysis and visuaLizAtion tool

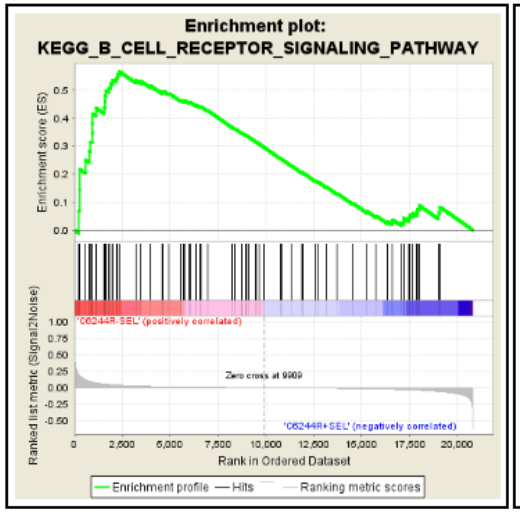
- Categorical or ranked analysis
- Mostly GO gene list support
- Interesting visualisation options

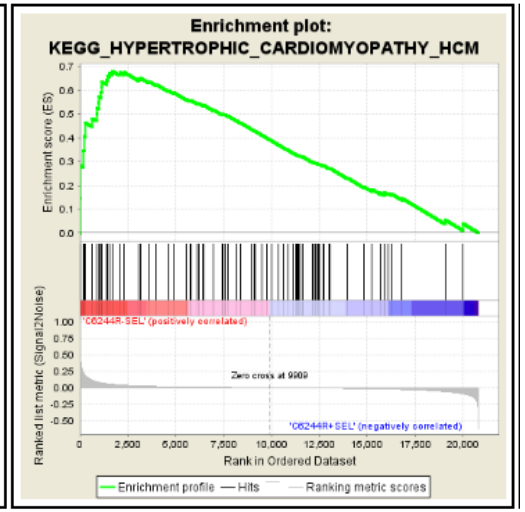


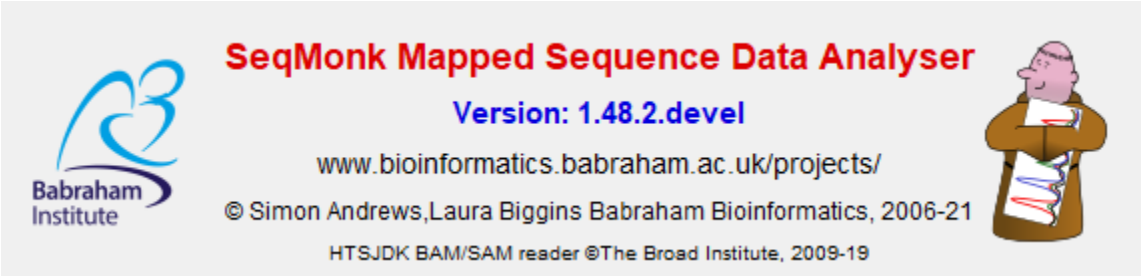
GSEA

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

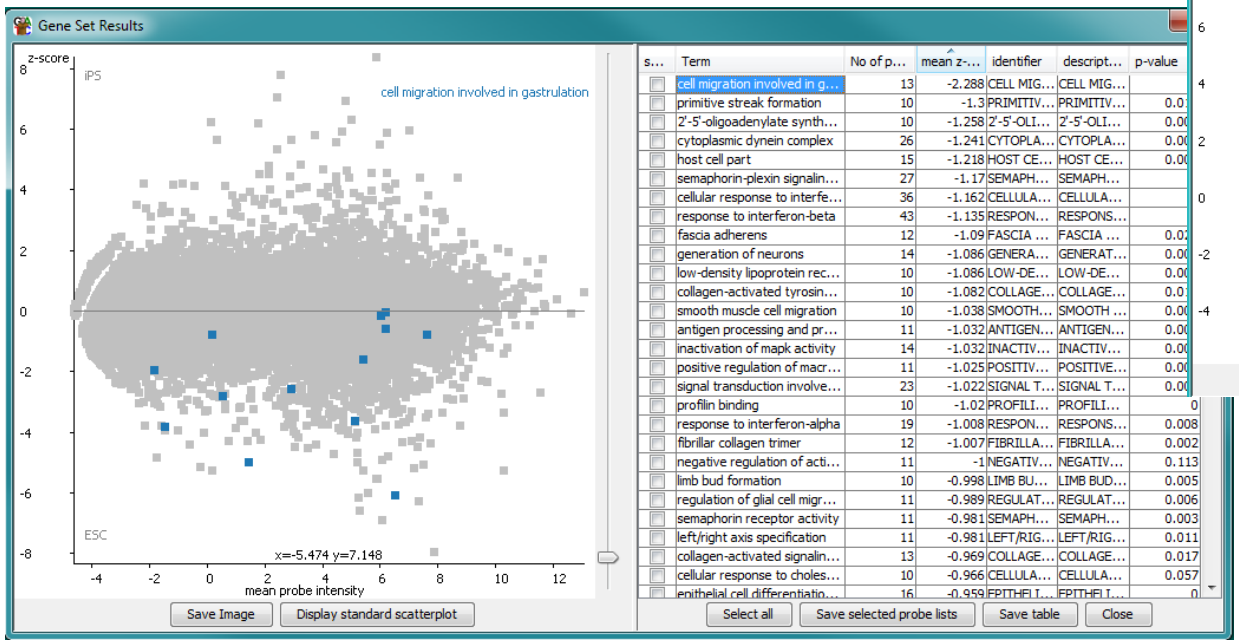


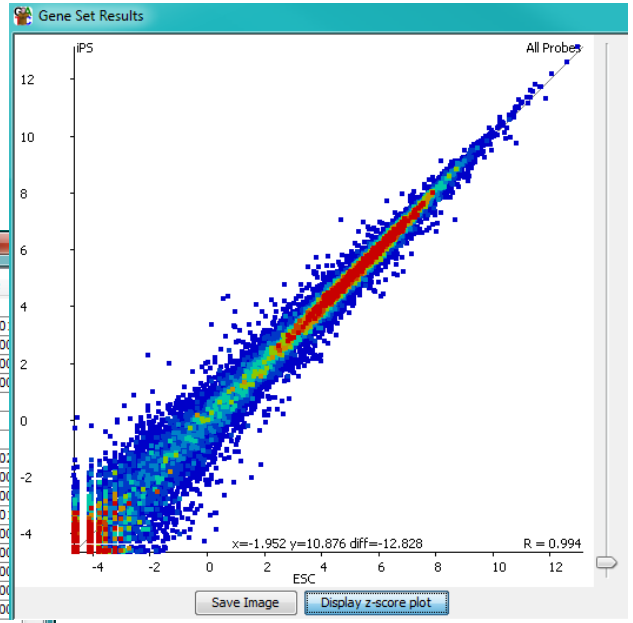






- Quantitative enrichment of sequencing datasets
- Local Java application





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

Babraham Bioinformatics

Functional enrichment results

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

Functional enrichment results

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

Tables are often enough

Category	*P* 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

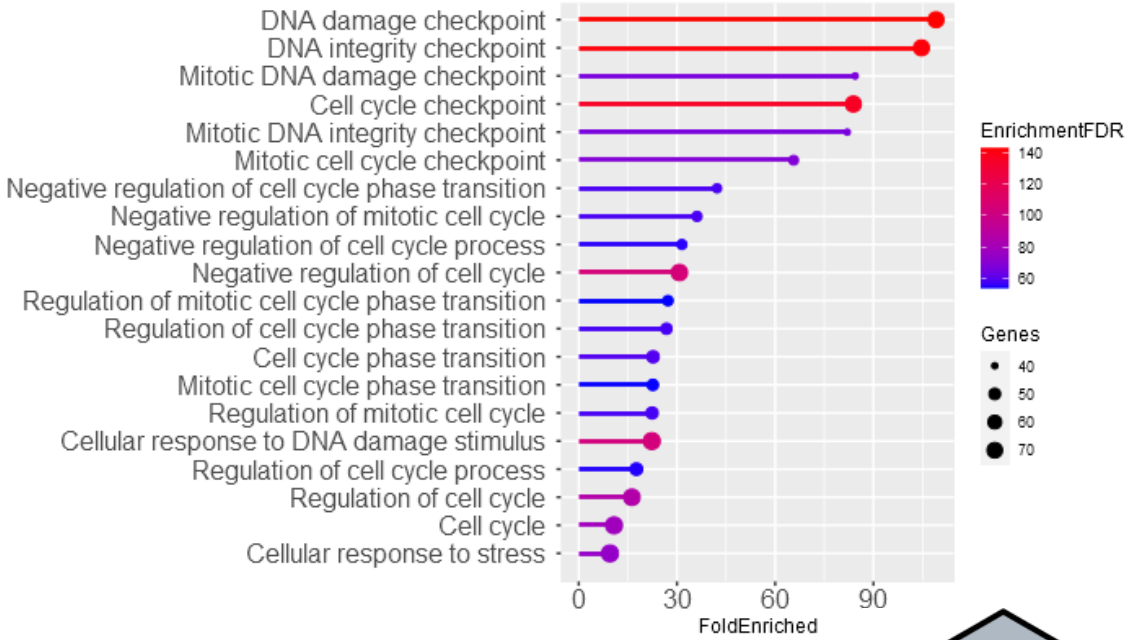
[1] percentage of cluster genes (relative to all genes on array) annotated for a given ontology term; [2] modified Fisher's Exact Test; [3] Benjamini and Hochberg false discovery rate-adjusted *p* value

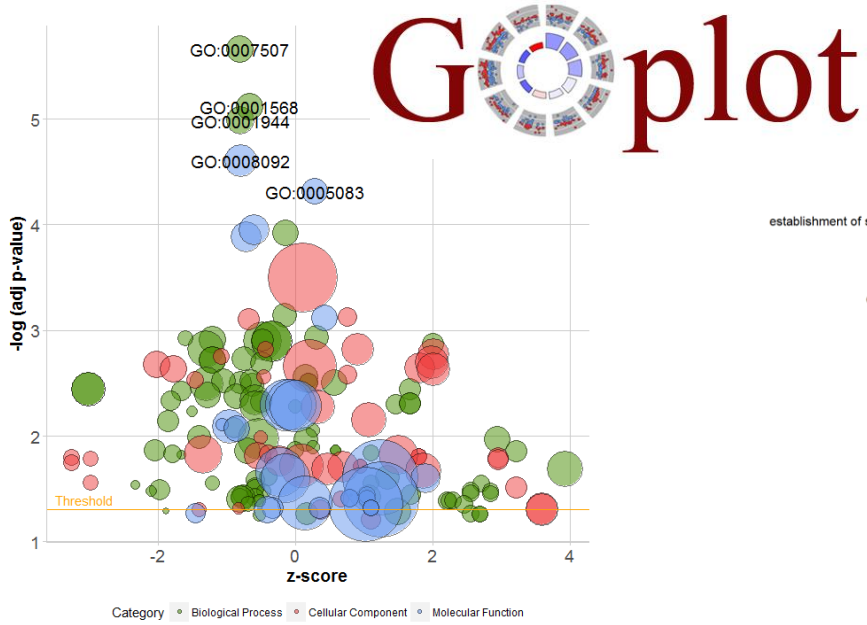
Graphical Representations

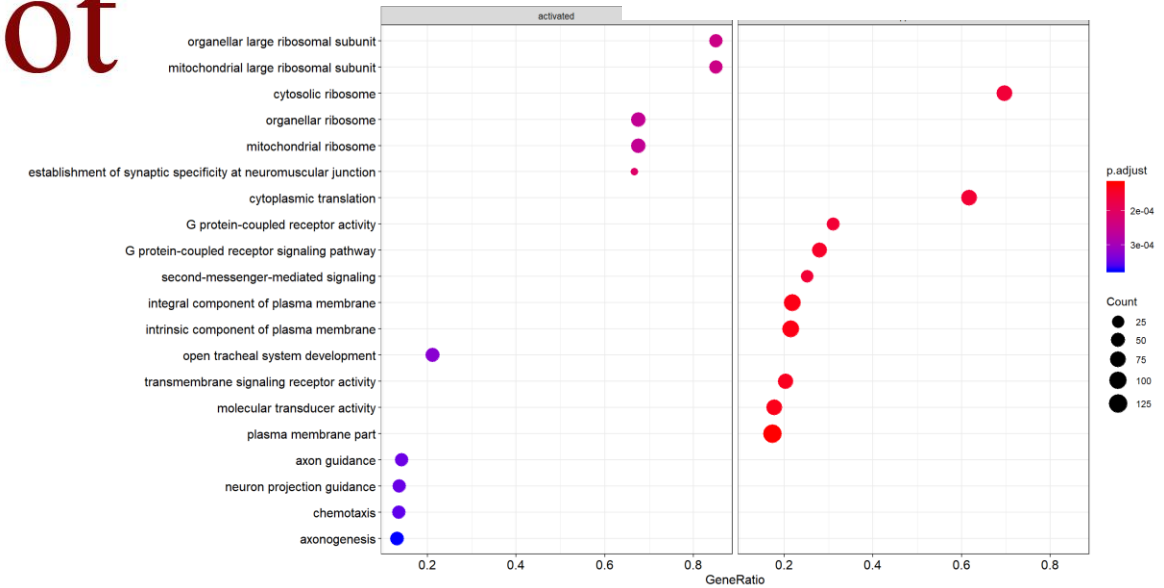
- Need to add something over a table
  - Relationships between multiple result values
  - Representation of redundancy between categories
  - Relationship to original data
  - Context of surrounding pathway

Plotting relationships between values

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

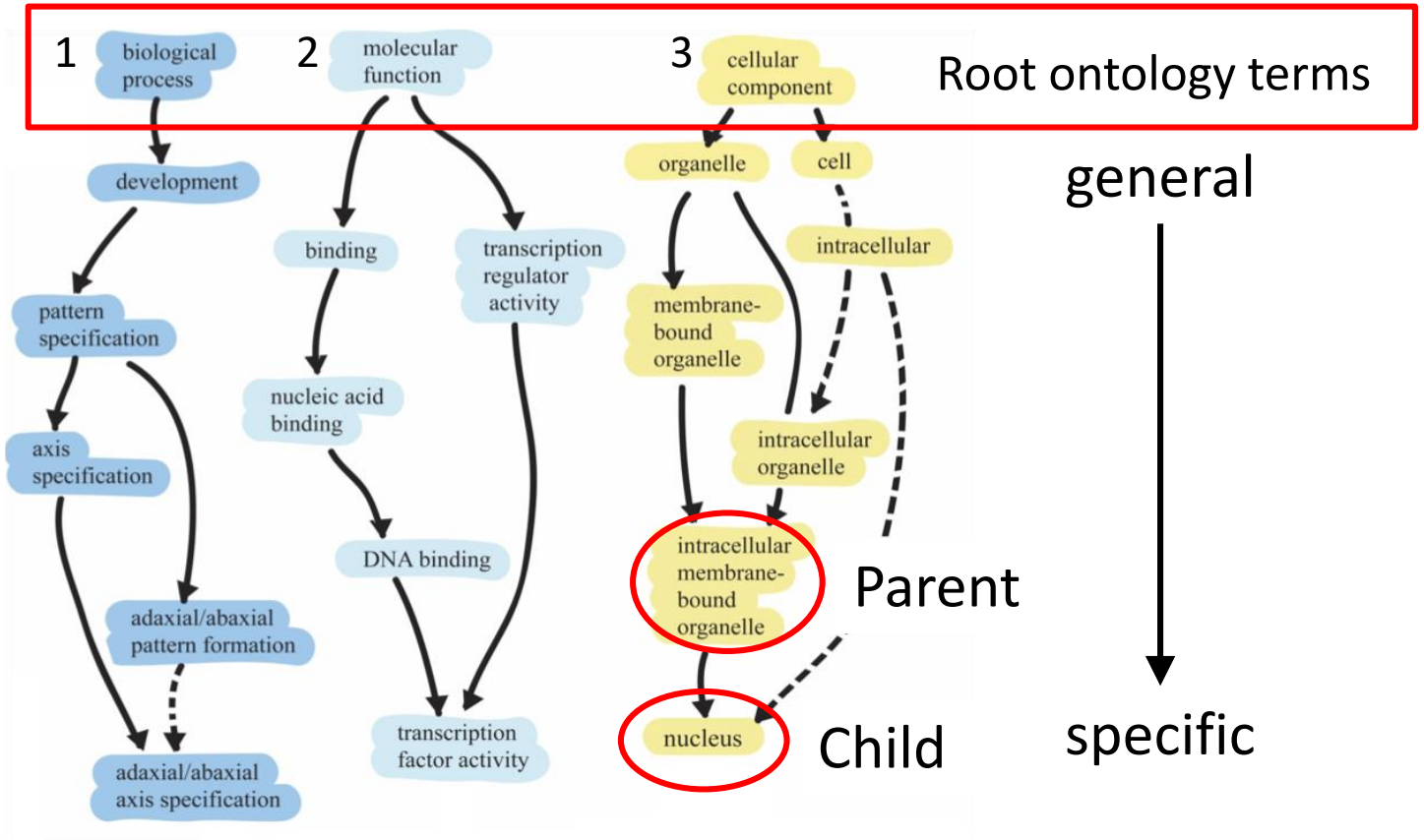






Redundancy in gene lists

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

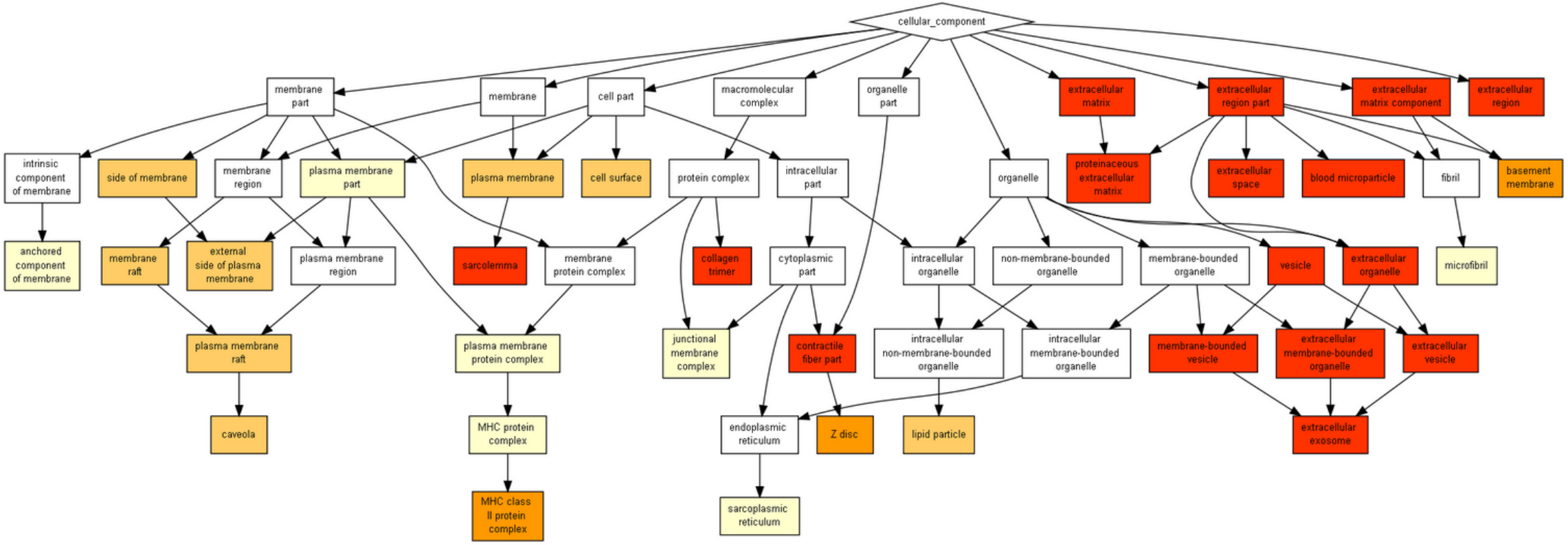


Redundancy: DAVID clustering

40 Cluster(s) Download File

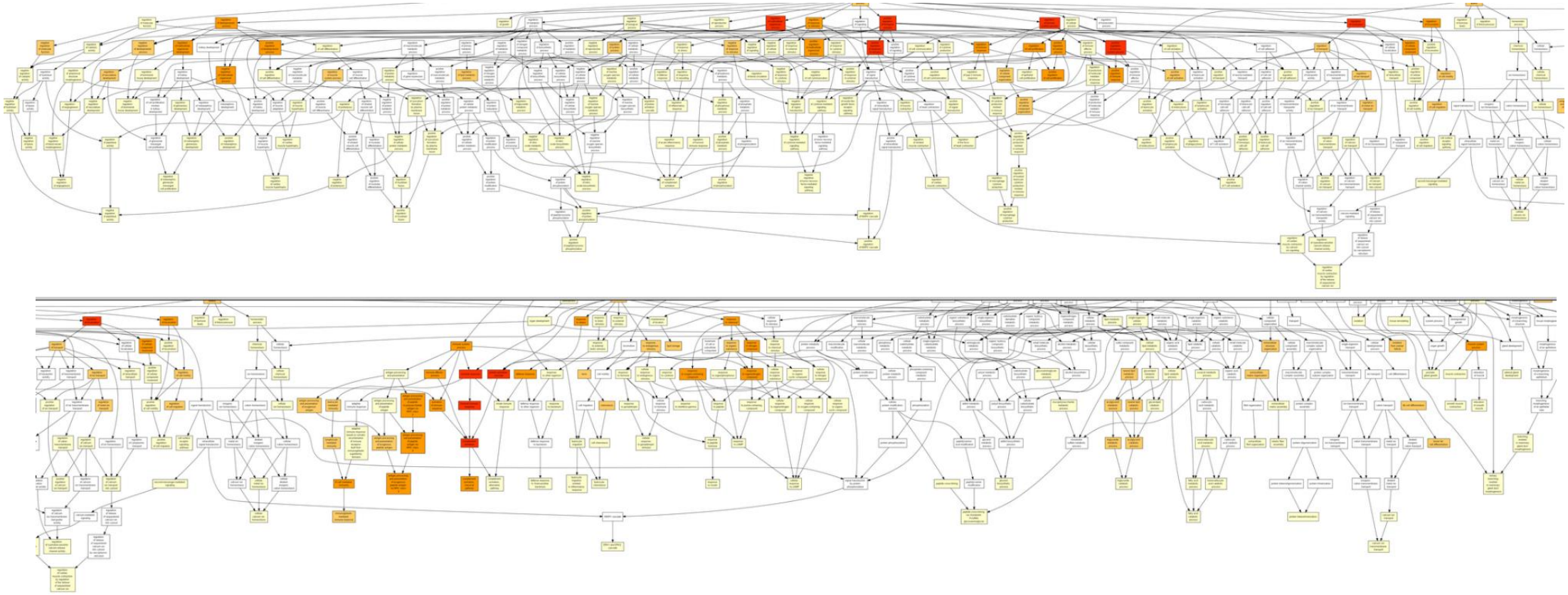
Annotation Cluster 1	Enrichment Score: 16.36		Count	P_Value	Benjamini
SP_PIR_KEYWORDS	dna-binding	RT	53	3.5E-24	4.5E-22
GOTERM_BP_FAT	regulation of transcription	RT	60	2.0E-20	1.8E-17
GOTERM_MF_FAT	DNA binding	RT	54	6.0E-20	7.9E-18
GOTERM_MF_FAT	transcription regulator activity	RT	45	2.0E-19	1.3E-17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

Redundancy: Gorilla GO images



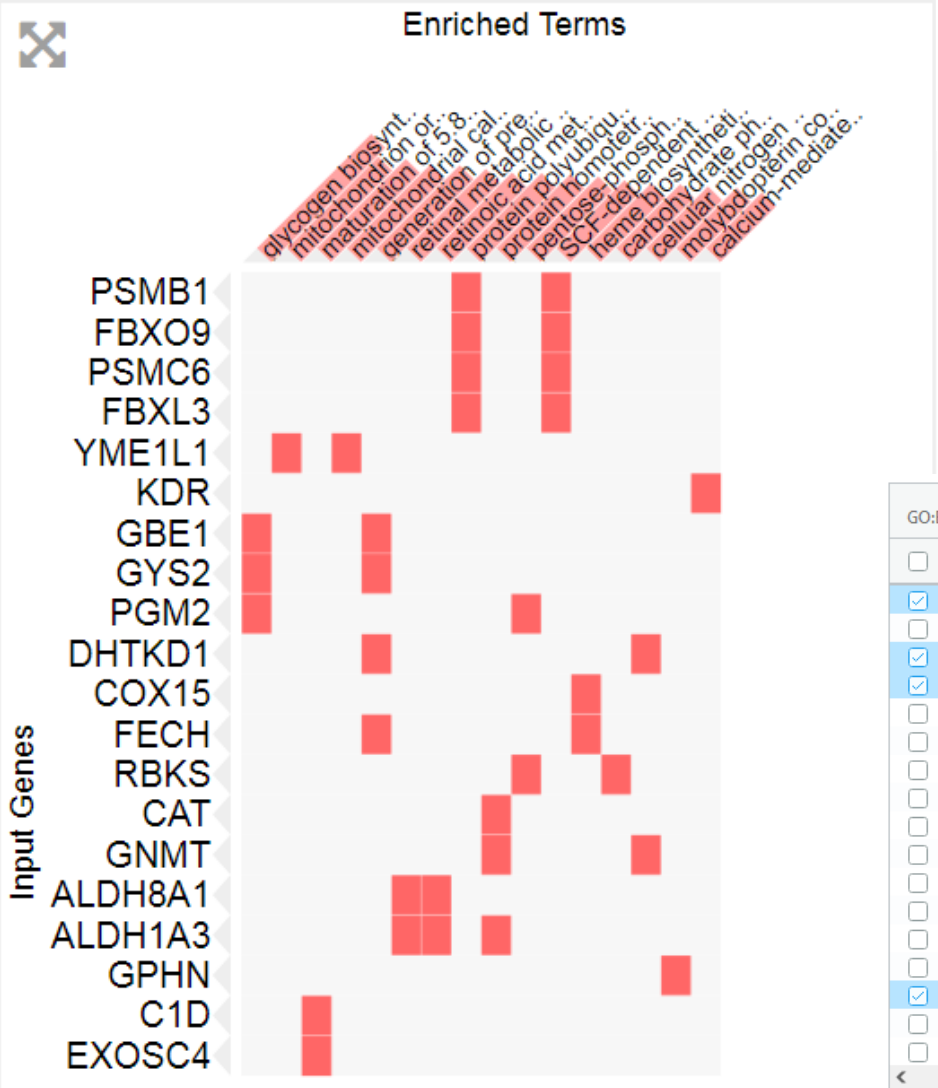
cbl-**gorilla**.cs.technion.ac.il/

Redundancy: Gorilla GO images



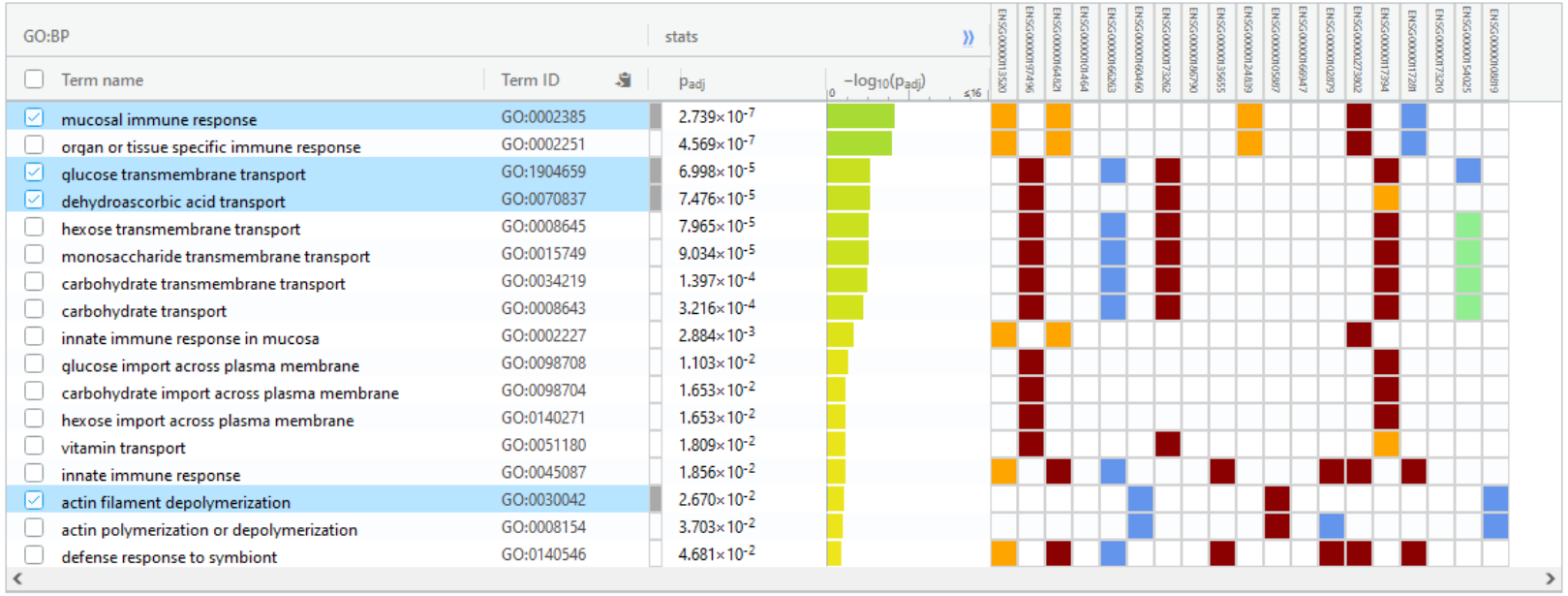
cbl-**gorilla**.cs.technion.ac.il/

Redundancy: Overlap plots



Enrichr

g:Profiler



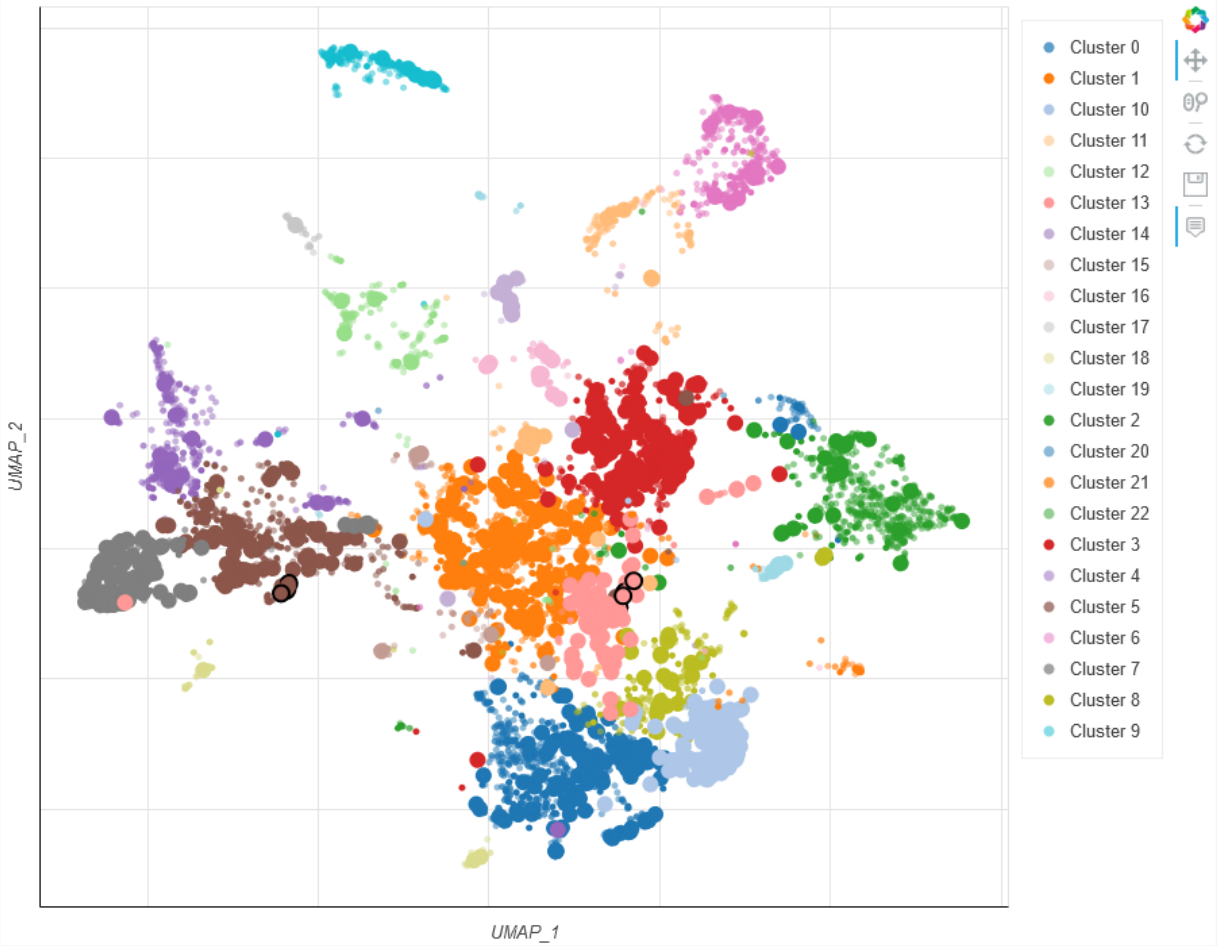
2D Redundancy

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

Enrichment Analysis Visualisation (from Enrichr)

https://appyters.maayanlab.cloud/Enrichment_Analysis_Visualizer/

Scatter plot visualization for GO_Biological_Process_2023.

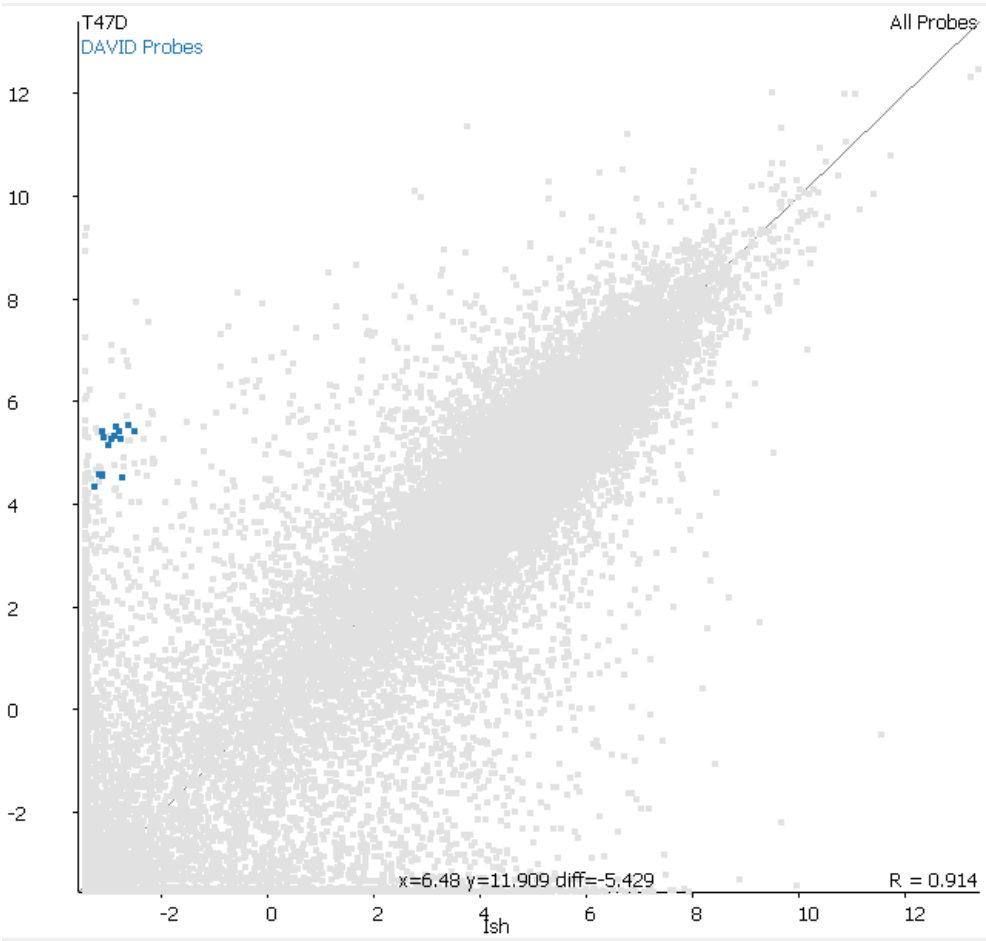


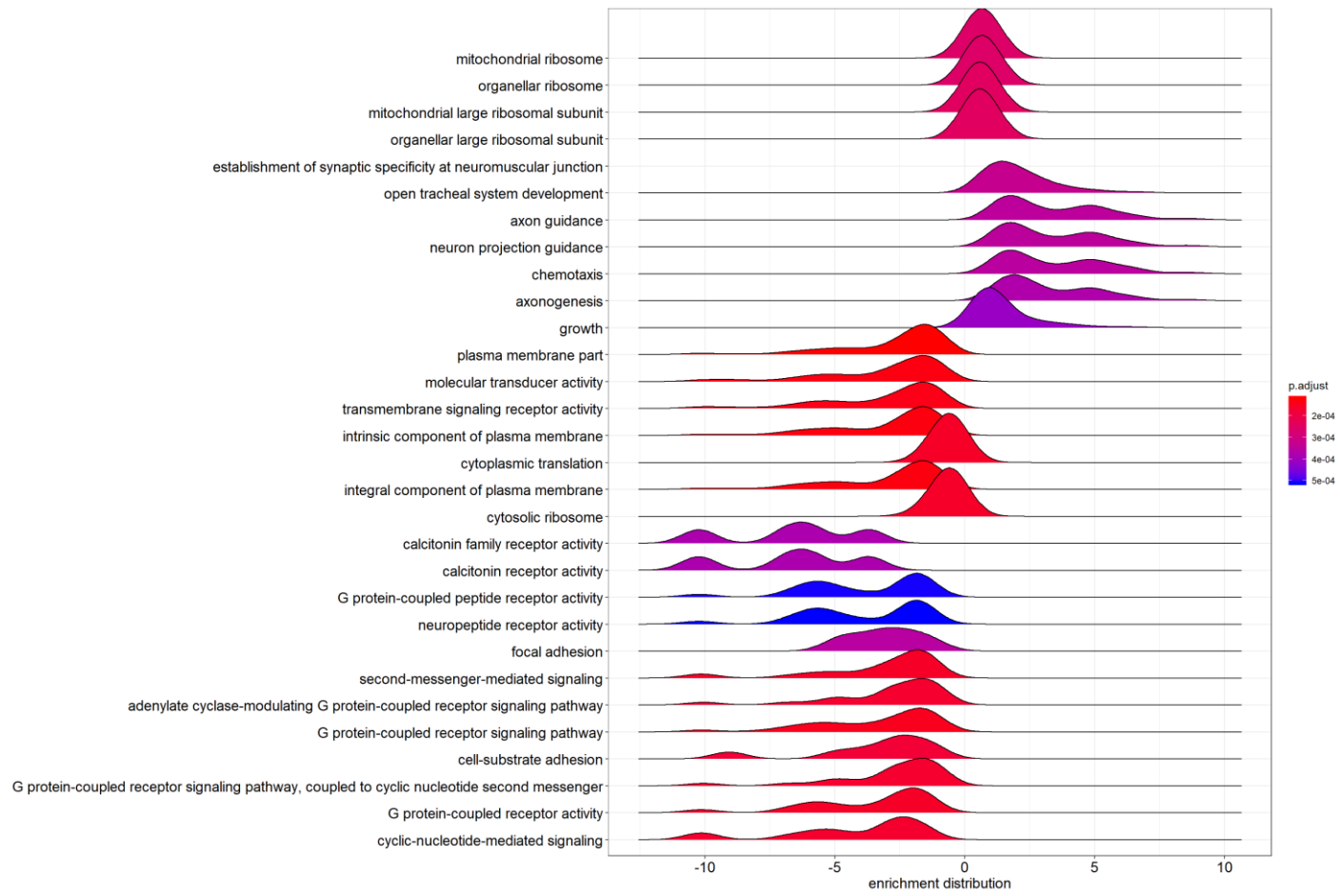
Relationship to original data

- Quantitative values for genes in category
  - Direction and magnitude of change

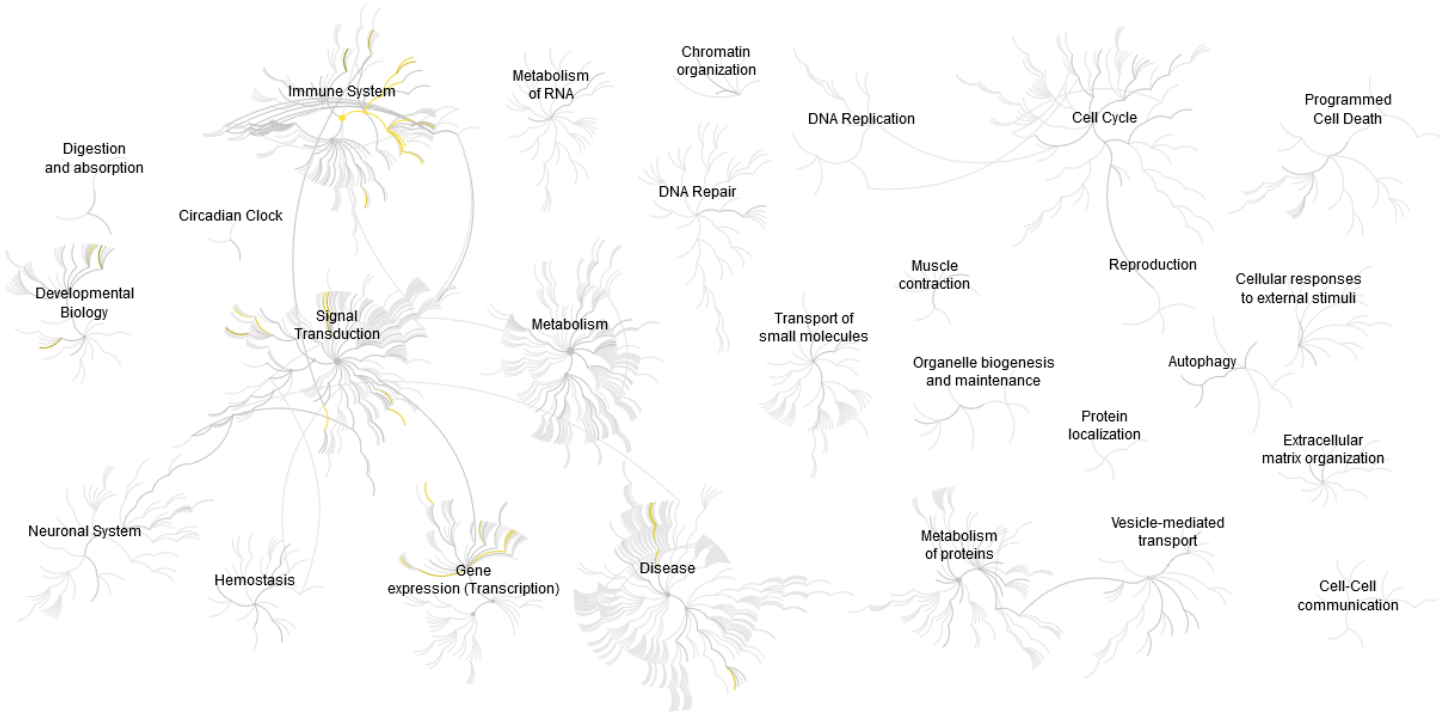
- Look at genes in category which aren’t hits
  - Relative numbers
  - Supportive changes?

Relationship to original data



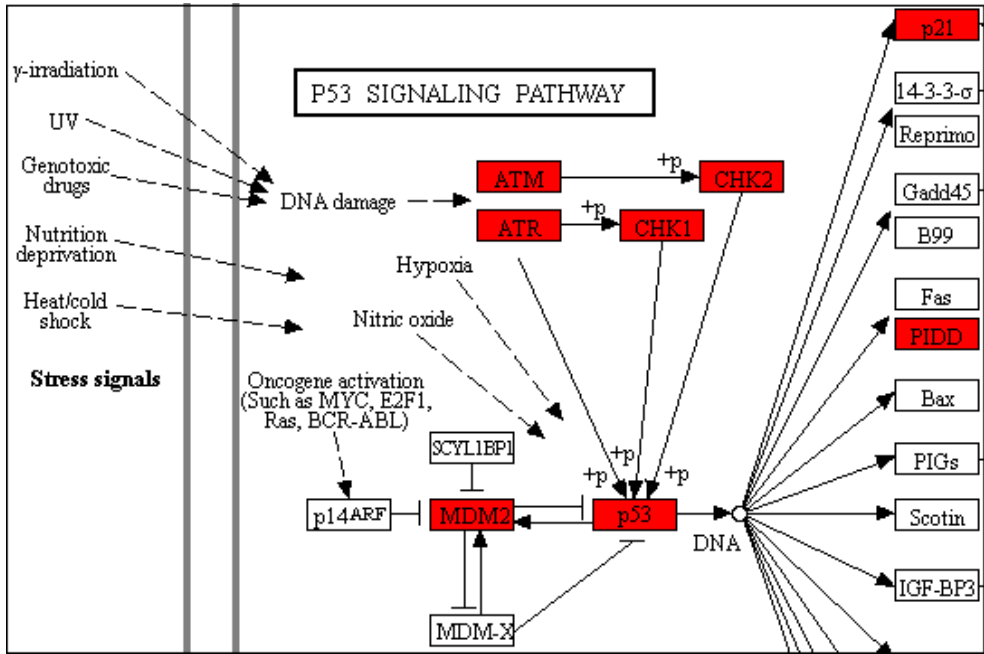


Pathways

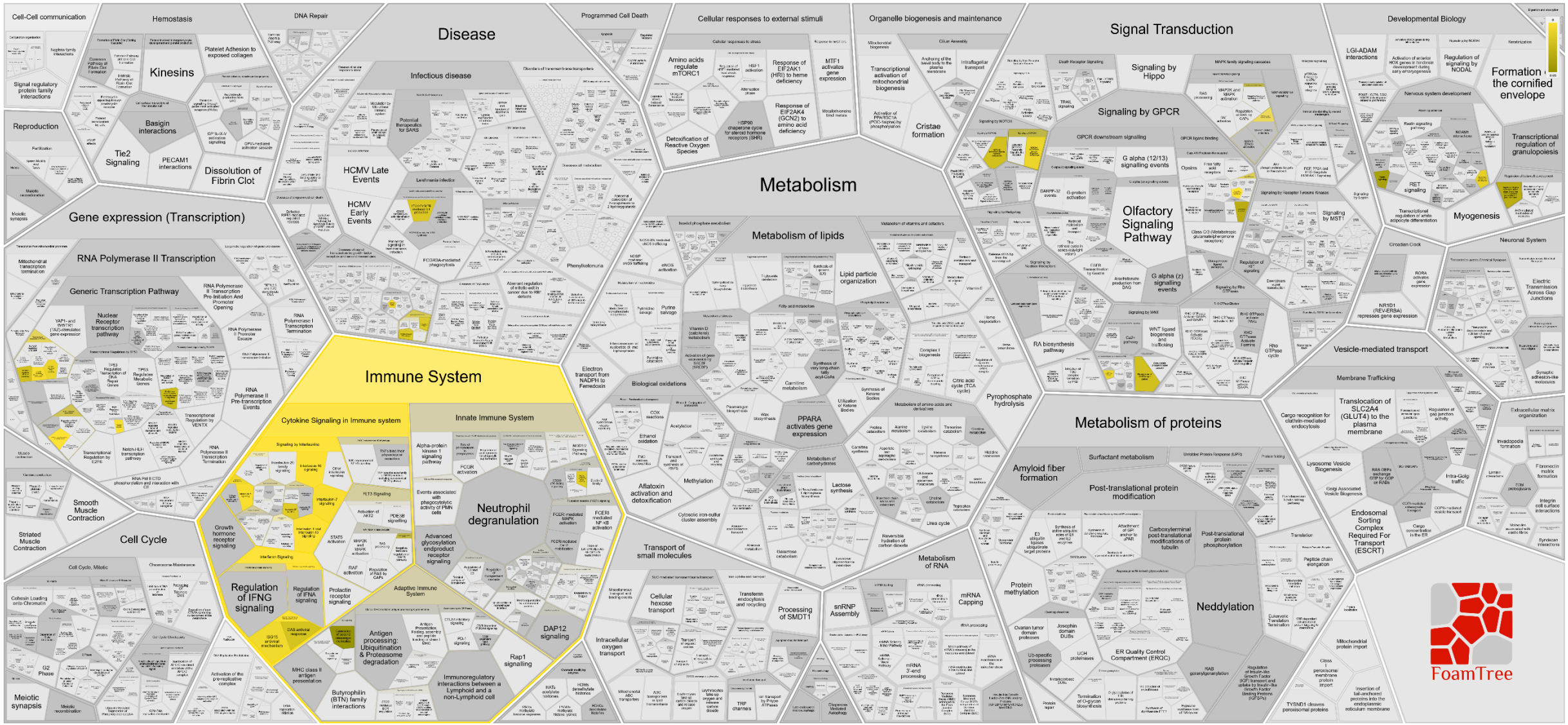


Reactome

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

Babraham Bioinformatics

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

Babraham Bioinformatics

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?

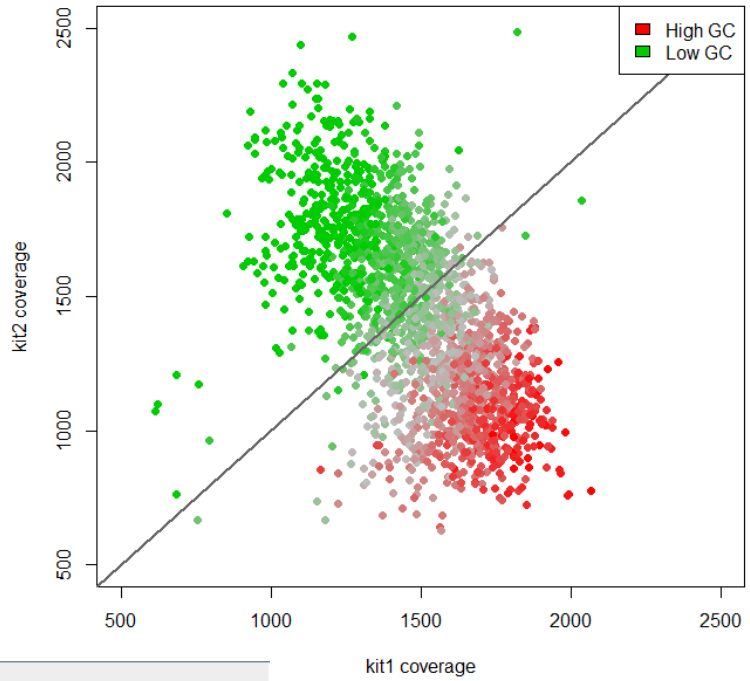
- Are some genes **never likely** to turn up in the hit list for technical reasons?

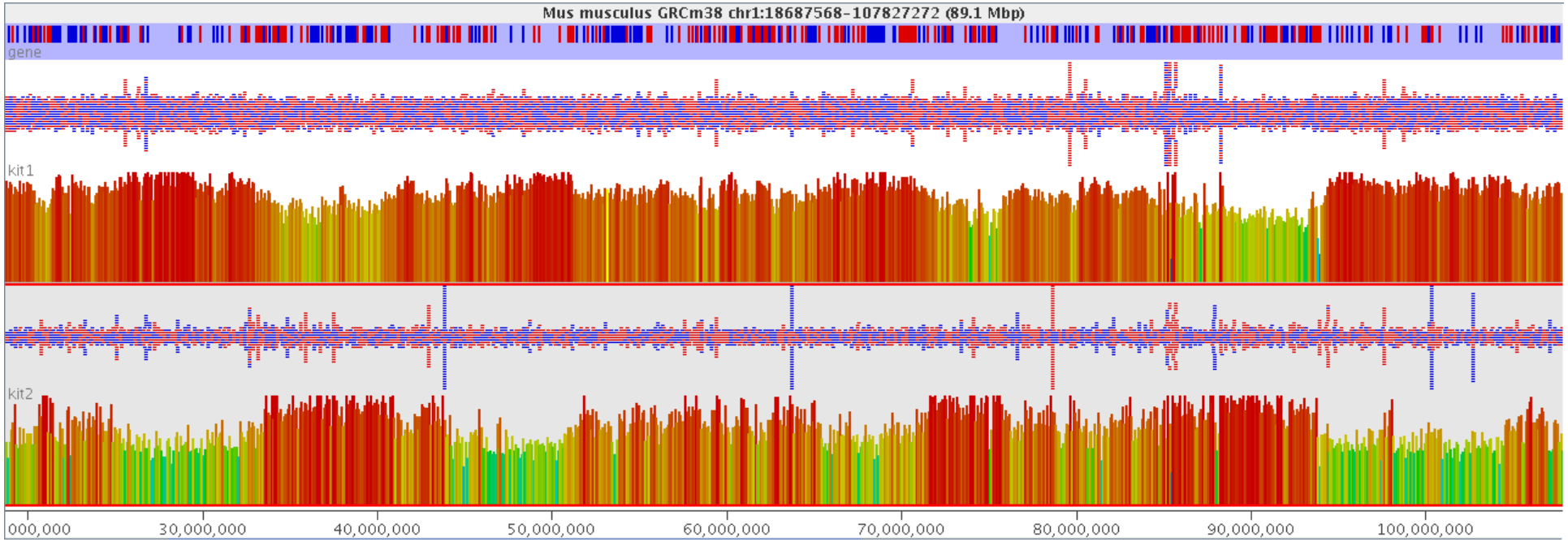
Biases

- 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)

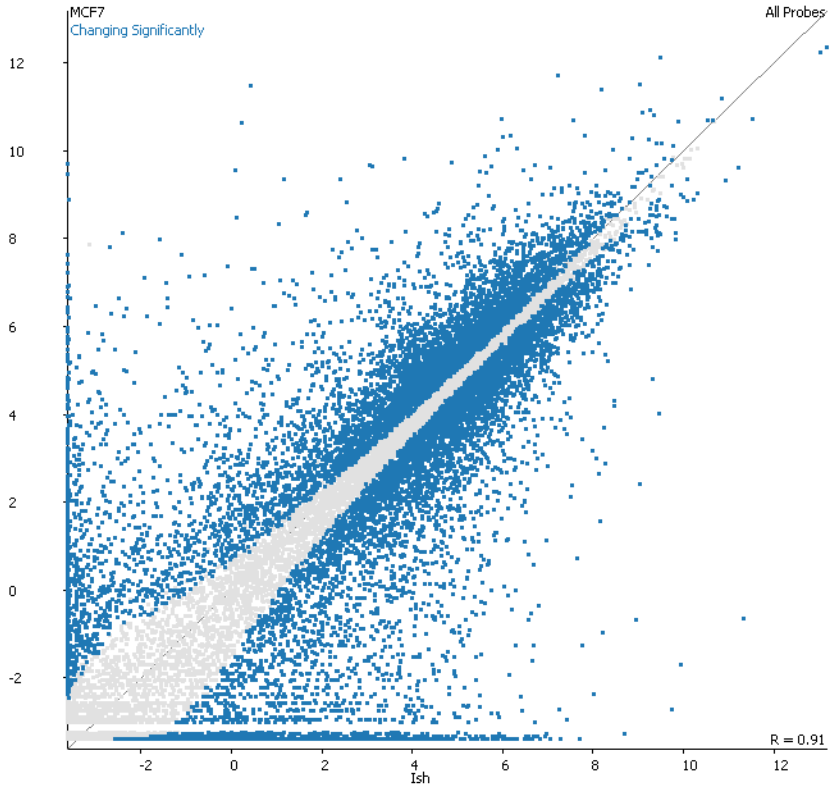
- Lists of hits are often biased based on statistical power

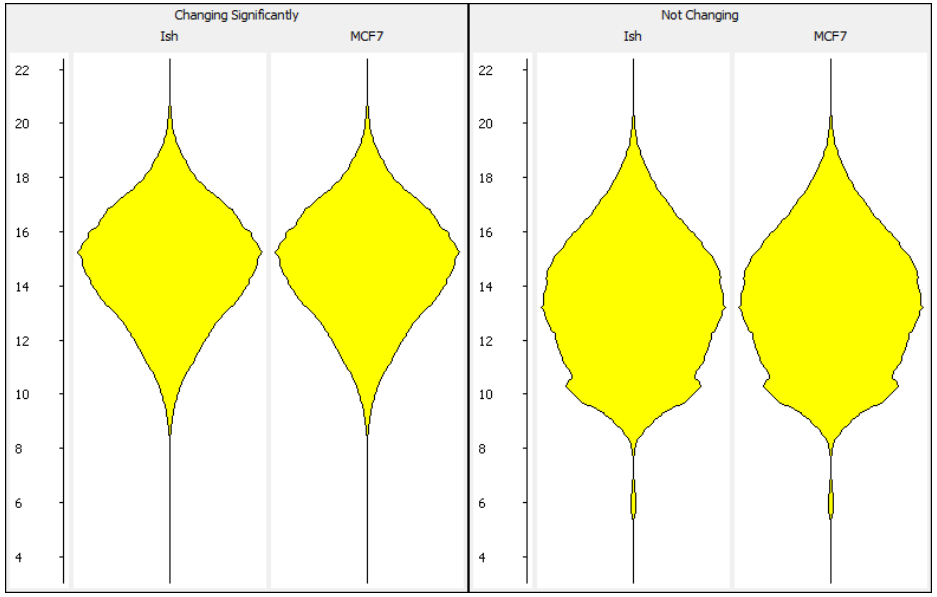
RNA-Seq Statistical Biases

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?

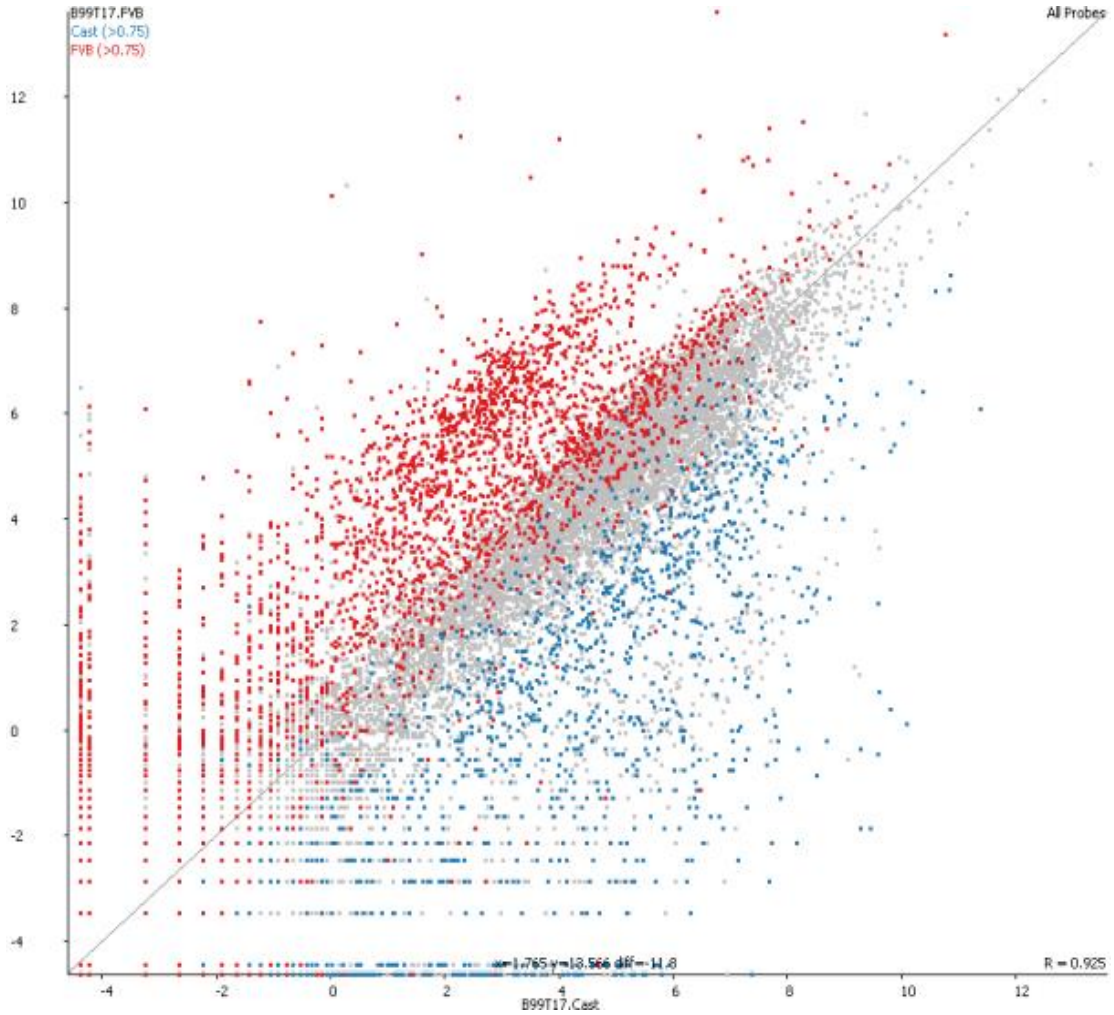
RNA-Seq Statistical Biases

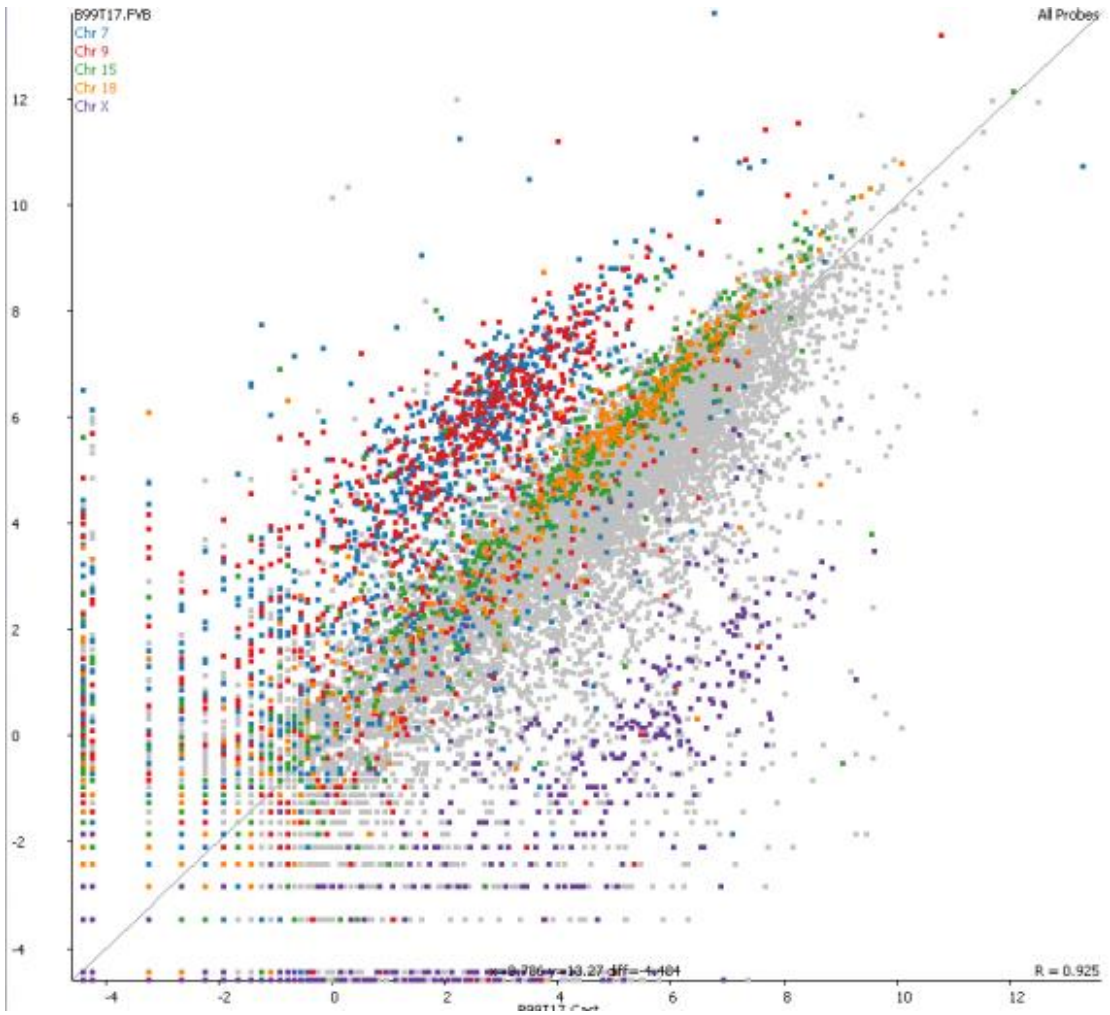




- Unlikely to ever see hits from genes which are
  - Lowly expressed
  - Short

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

Hindawi Publishing Corporation
Evidence-Based Complementary and Alternative Medicine
Volume 2016, Article ID 7276161, 13 pages
http://dx.doi.org/10.1155/2016/7276161

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]

[1]*Shanghai 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*
[2]*Tumor Institute of Traditional Chinese Medicine, Longhua Hospital, Shanghai University of Traditional Chinese Medicine, 725 South Wanping Road, Shanghai 200032, China*
[3]*College 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.

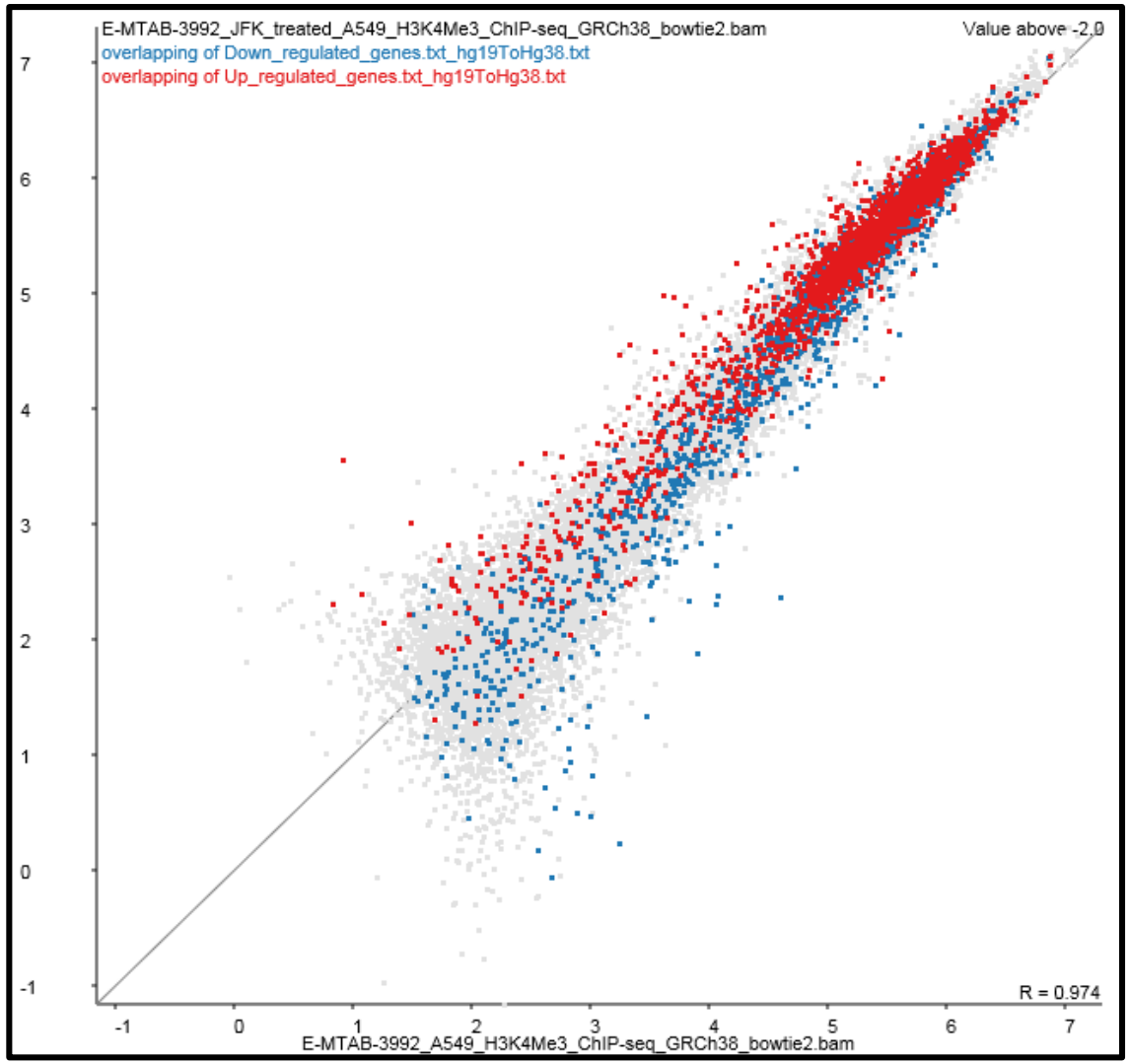
Traditional Chinese medicine Jinfukang (JFK) has been clinically used for treating lung cancer. To examine whether epigenetic modifications are involved in its anticancer activity, we performed a global profiling analysis of H3K4Me3, an epigenomic marker associated with active gene expression, in JFK-treated 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

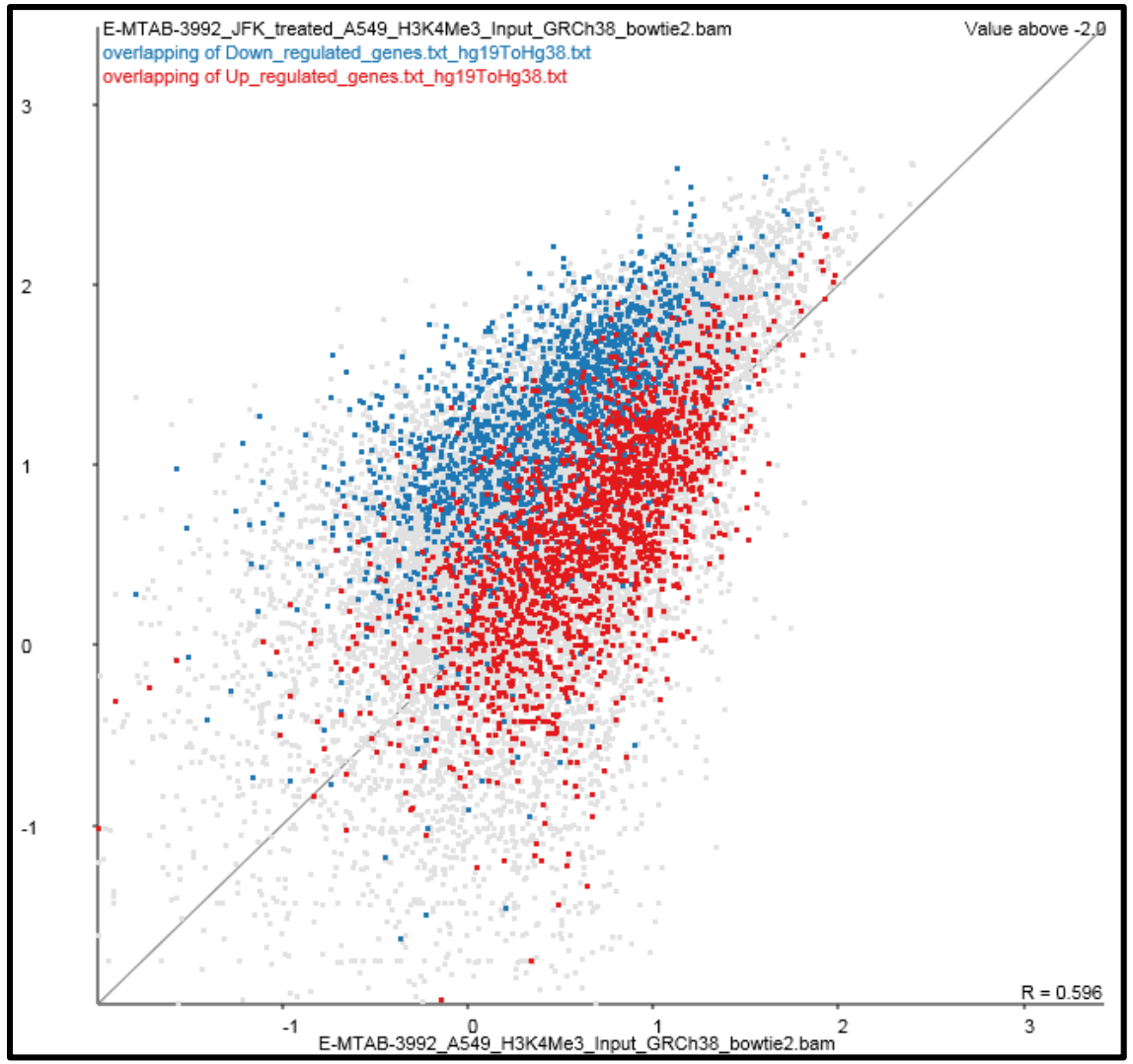
Chromatin is the macromolecular complex of DNA and histone proteins that provides the scaffold for packaging the eukaryotic genome [1, 2]. Histones H2A, H2B, H3, and H4 are the basic components of nucleosomes, which form the fundamental unit of chromatin [3, 4]. Chemical modifications to the histones 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

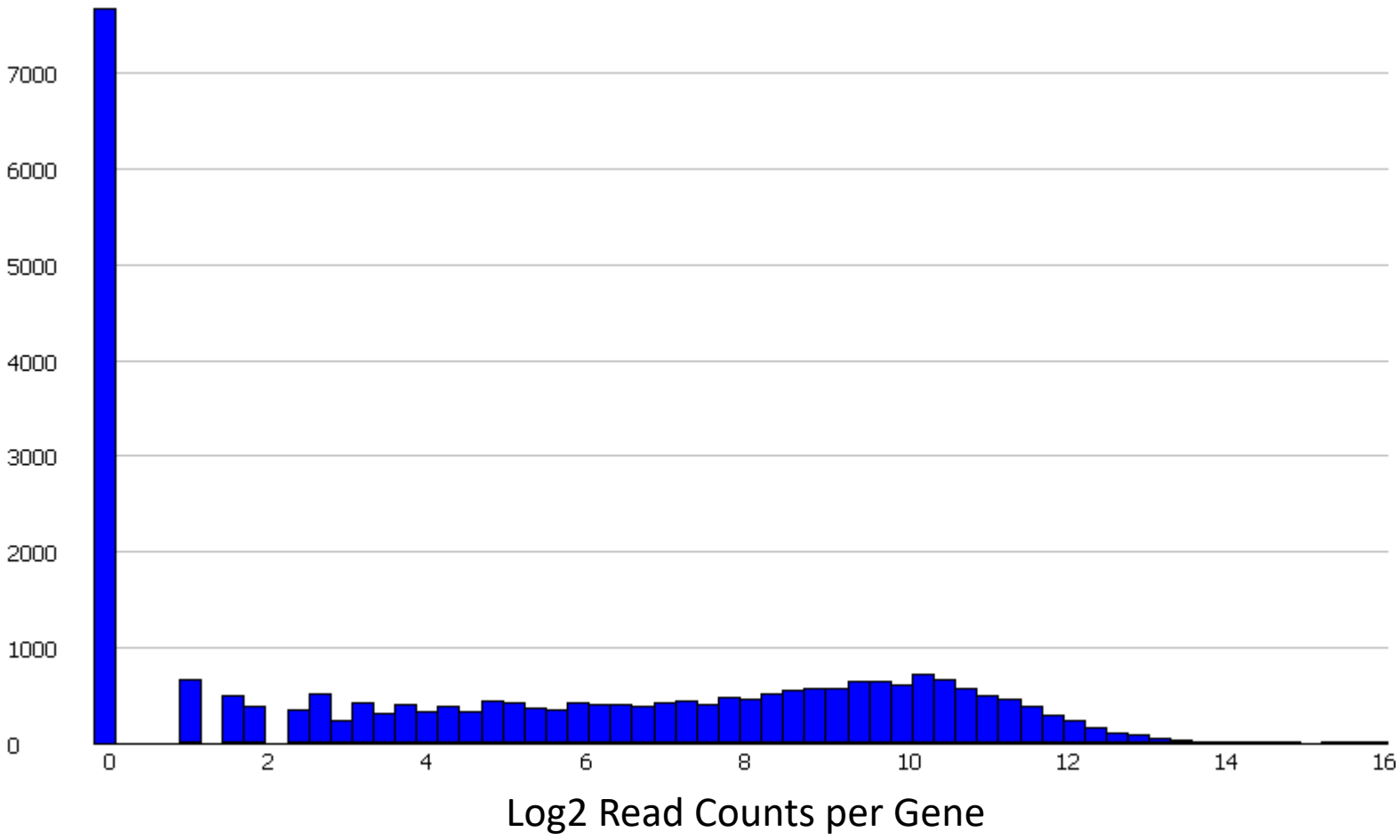
- Use custom backgrounds during Gene Set Analysis to help minimise bias (if a tool supports it)

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

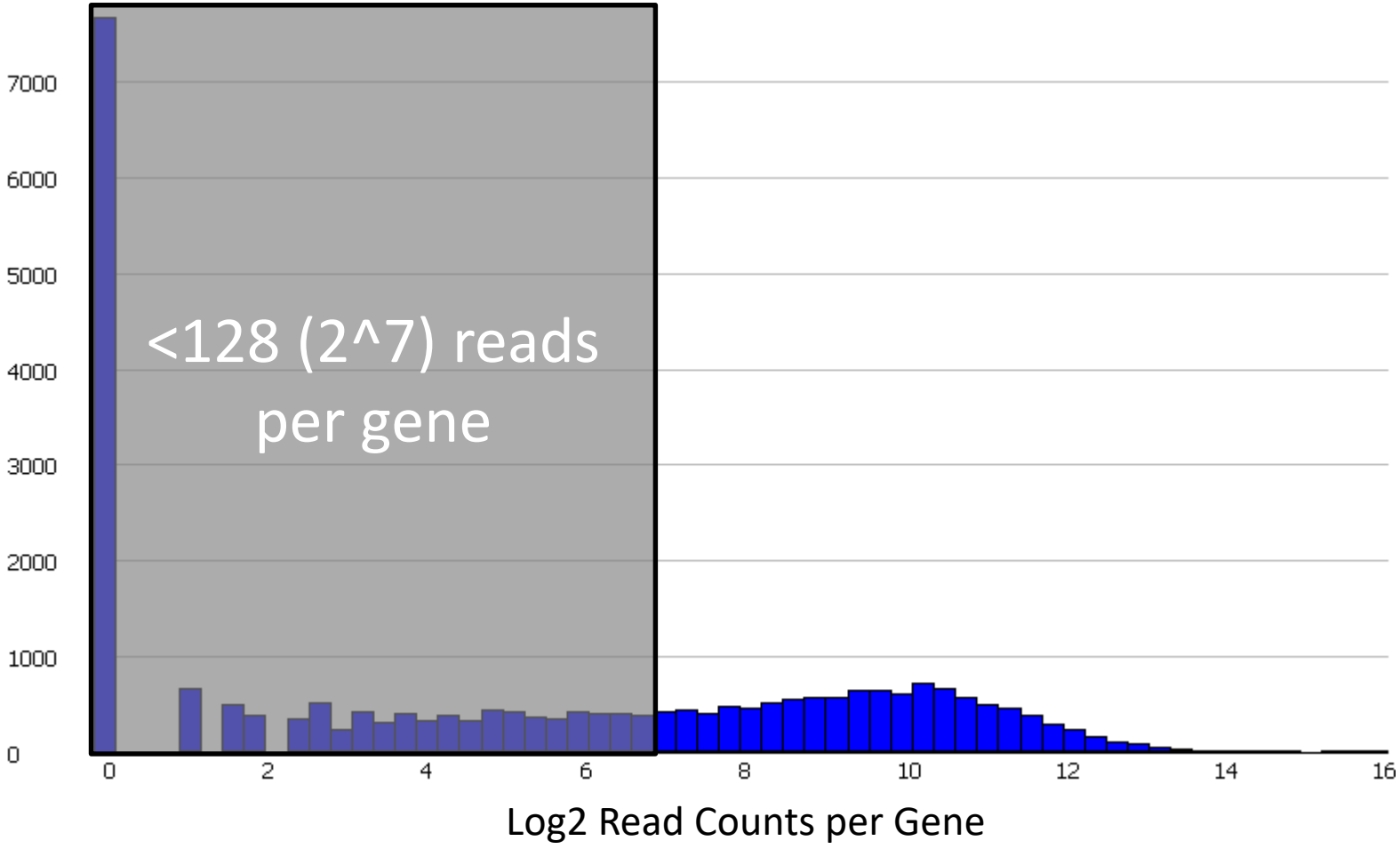
- Make a list of what you ***could*** have seen, and set that as the background.

Expressed Genes



26,127 Genes Measured

Expressed Genes



10,378 Genes Realistically Measured (40%)

Statistical biases affect gene sets too

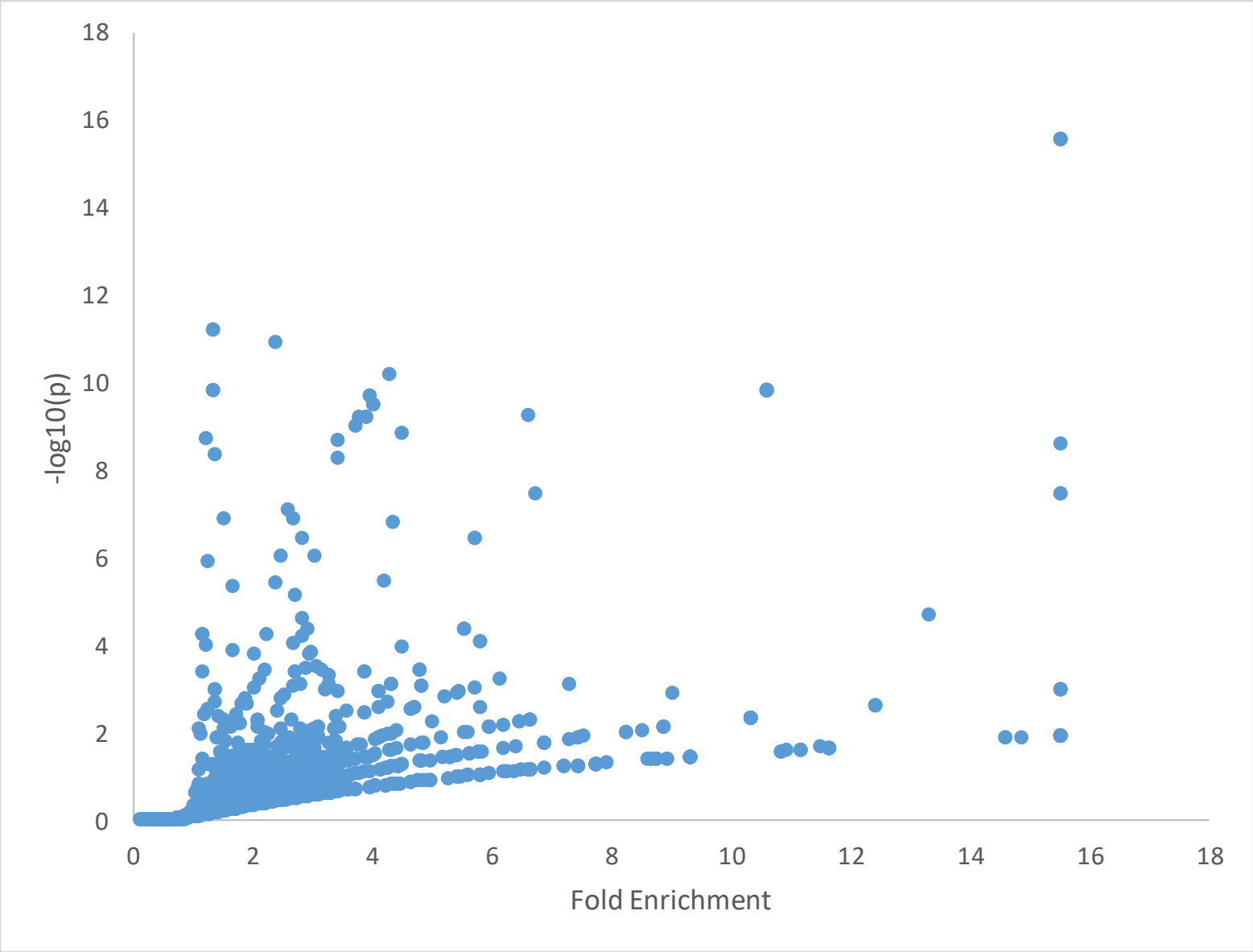
- Fisher's test is powered by
  - Magnitude of change
  - Observation level

- 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

GOliath

www.bioinformatics.babraham.ac.uk/goliath/

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

Welcome to GOliath

Select species	Homo_Sapiens/Dec_18
Min Category Size	50
Max Category Size	500
Gene List	Background List (optional)
Paste Gene Names here	Paste Gene Names here

Query name (optional)

Use example genes

Analyse my list

GOliath Gene Ontology Search System (Version 0.1.devel)

Having problems with the site? Please let Simon Andrews know

GOliath

www.bioinformatics.babraham.ac.uk/goliath/

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

Results for job Course test

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	

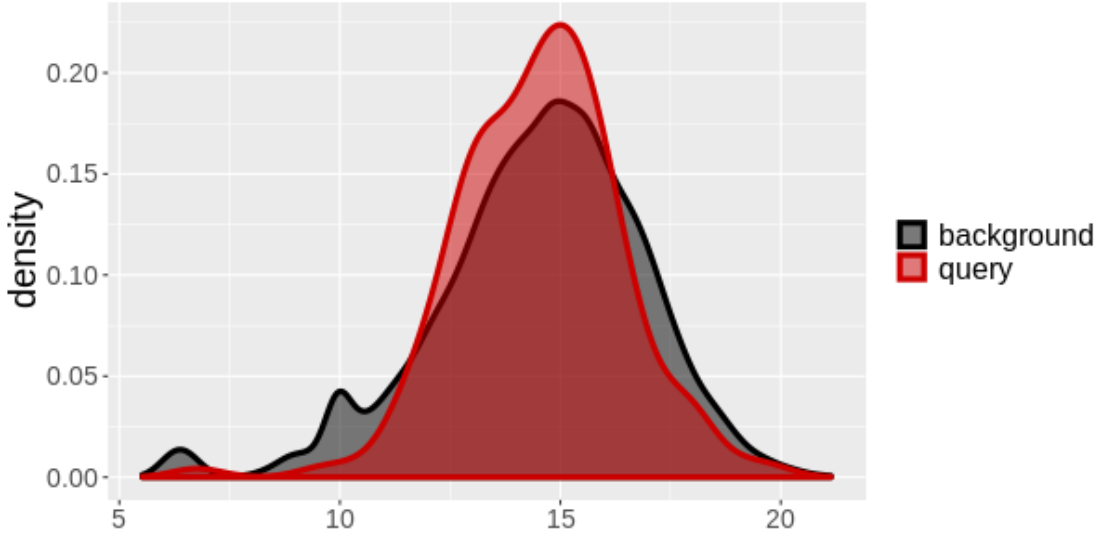
Checking for unexpected biases

- 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

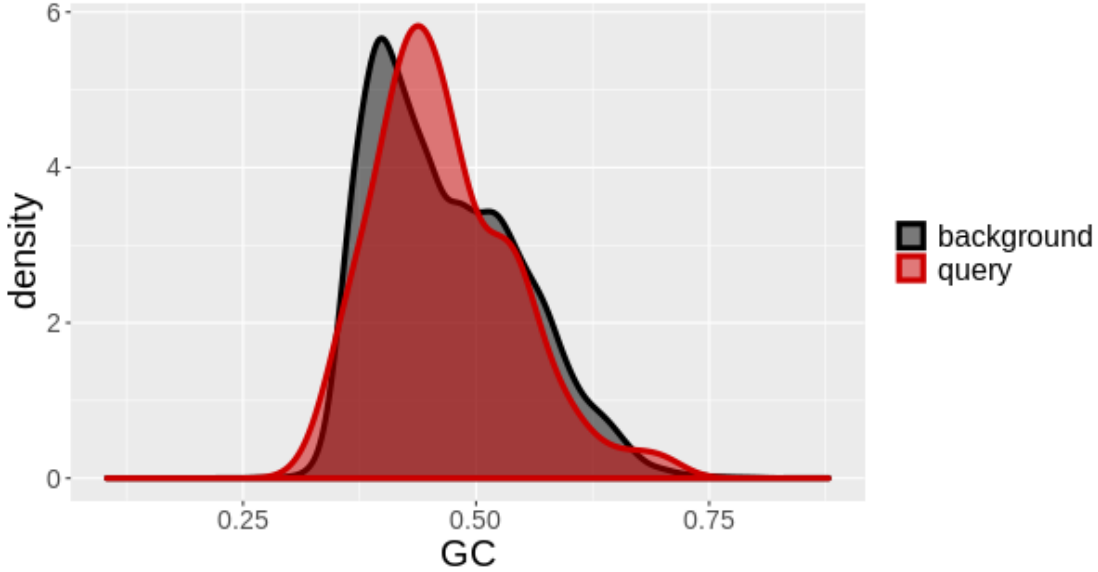
- If a bias exists then is this the actual link between genes? If not then can I fix this by improving my background list?

www.bioinformatics.babraham.ac.uk/goliath/

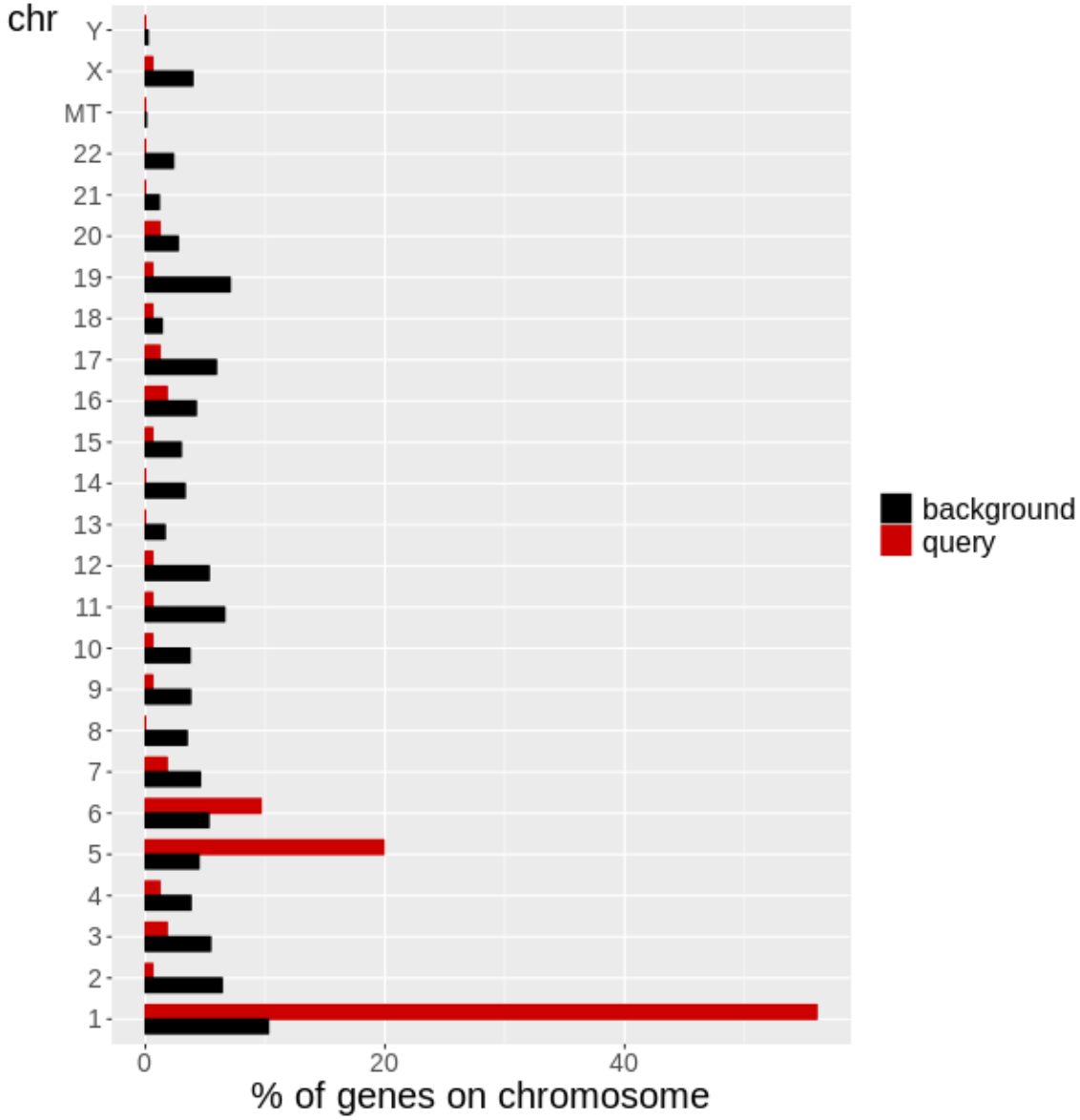
Gene lengths



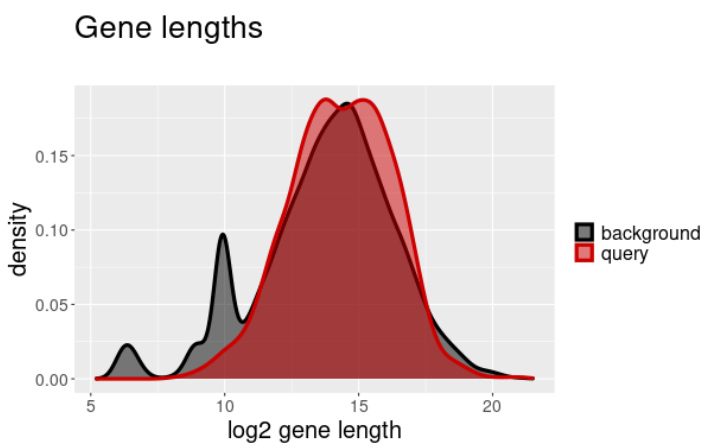
GC content of genes

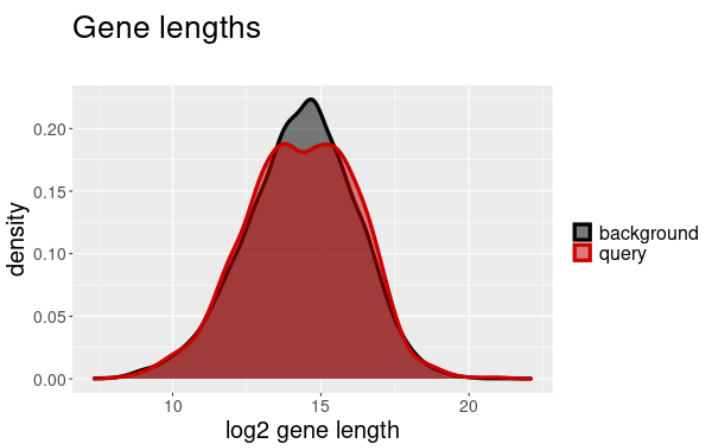


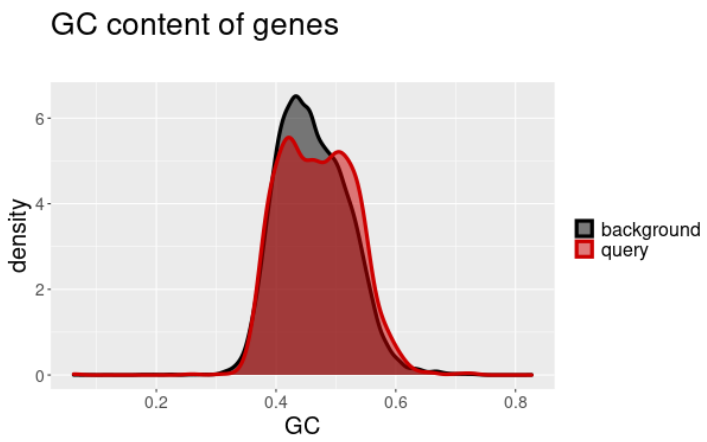
Chromosomal locations



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

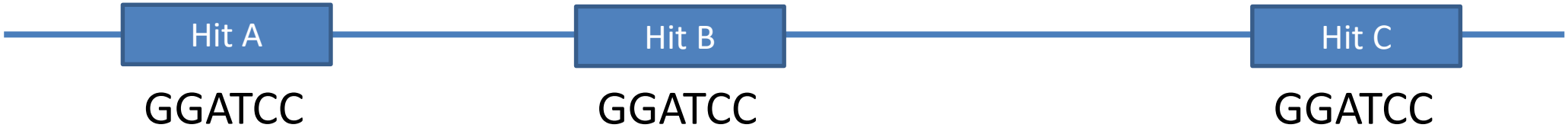
Motif Searching

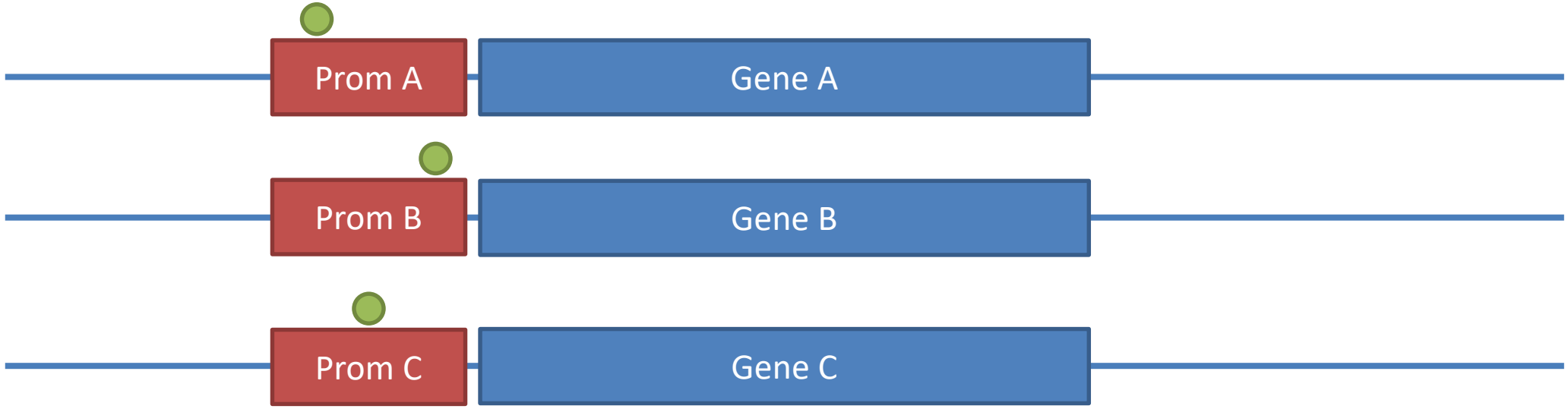
Simon Andrews

simon.andrews@babraham.ac.uk

Babraham Bioinformatics

Rationale

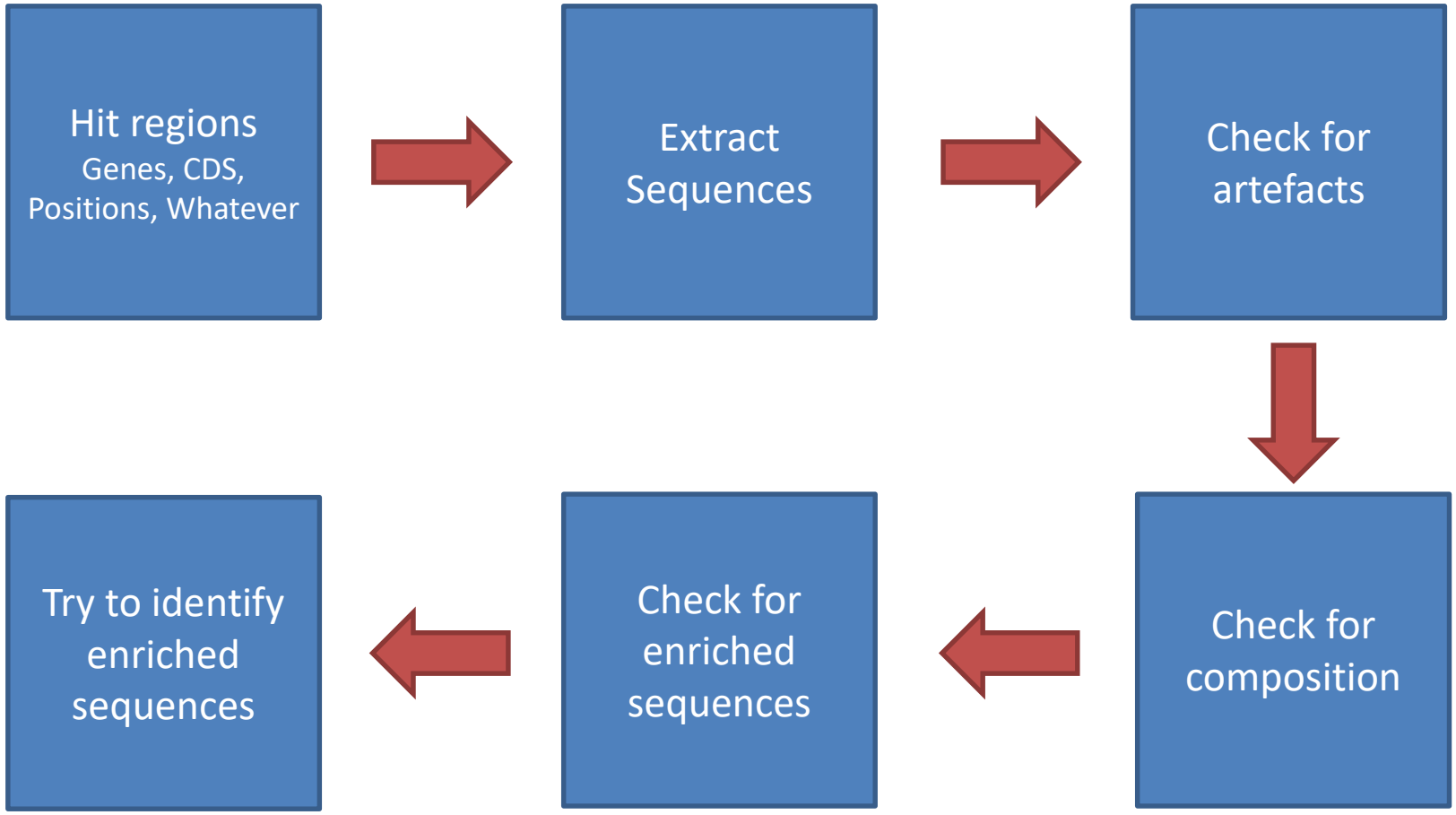

Hit A
Hit B
Hit C
GGATCC
GGATCC
GGATCC


Prom A
Gene A
Prom B
Gene B
Prom C
Gene C

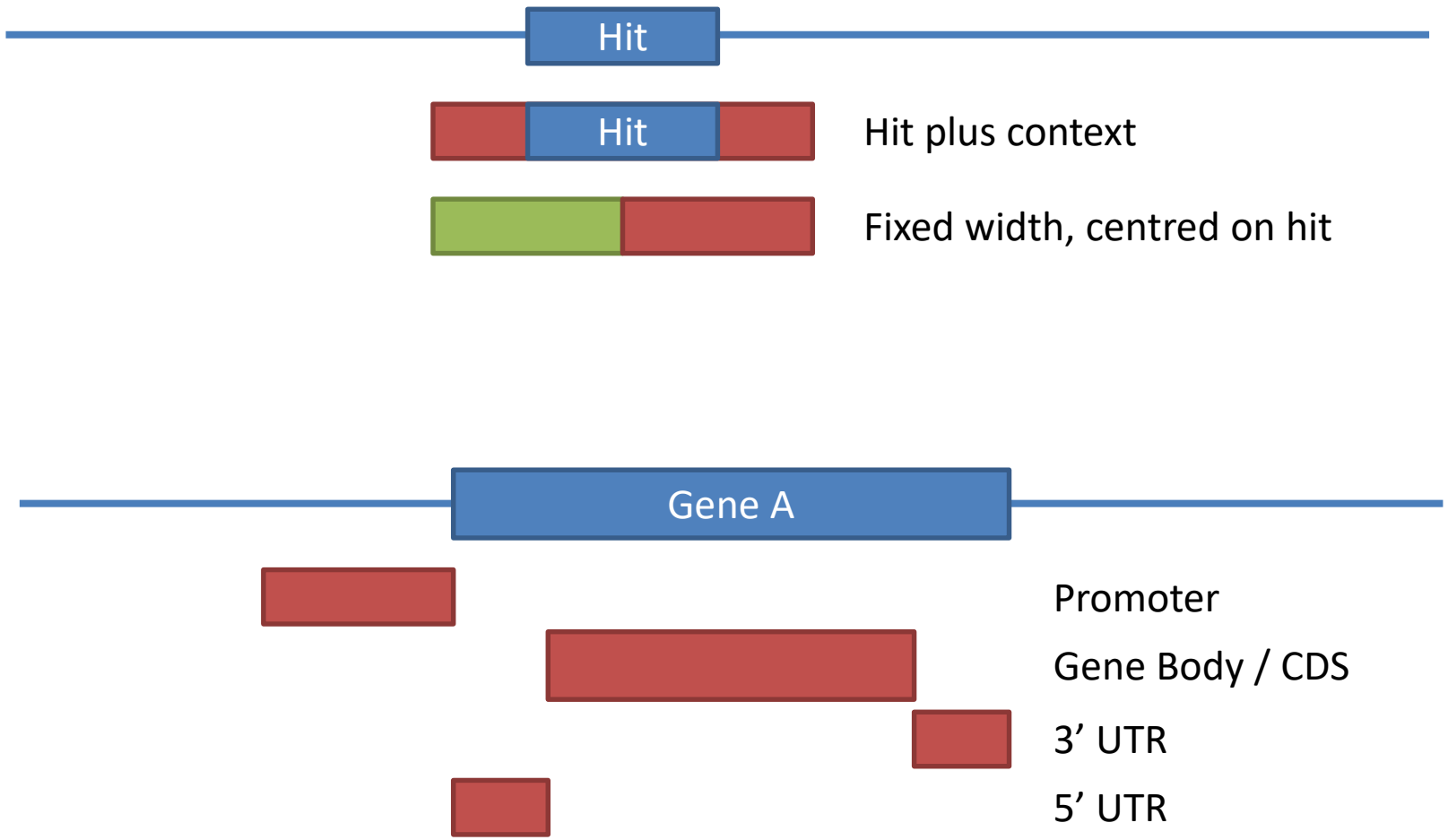
Basic Questions

- Does the sequence around my hits look unusual?
- Do specific sequences turn up more often than expected in my hits?
- If so, do the sequences look like any known functional sequence?

Basic Workflow



Deciding what to extract



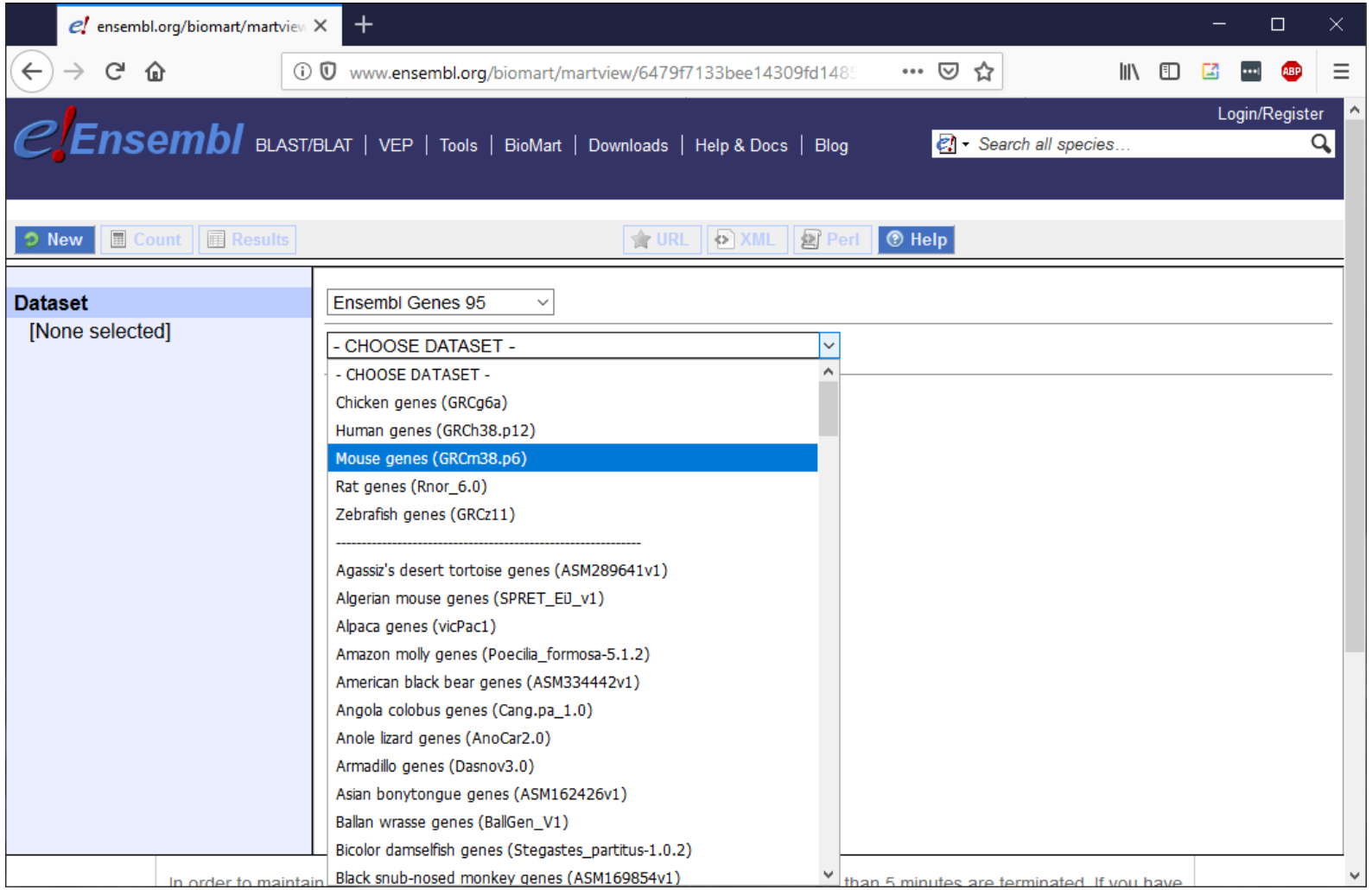
Extracting Sequence

- From positions
  - BEDTools
  - Genome Browsers*
  - Custom scripts

- From features
  - Genome Browsers*
  - BioMart

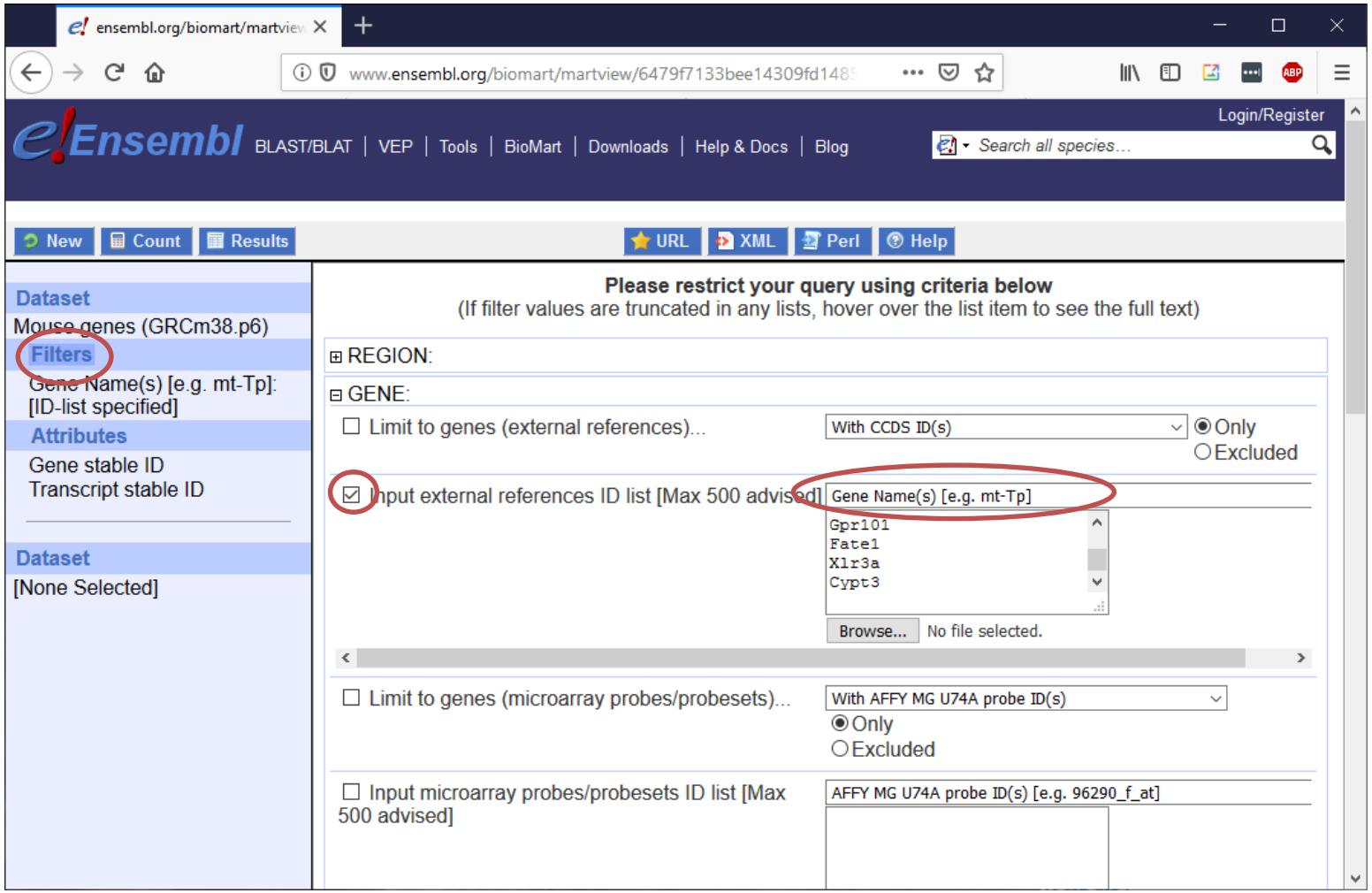
*not easily automatable for multiple sequences

BioMart – Selecting Assembly

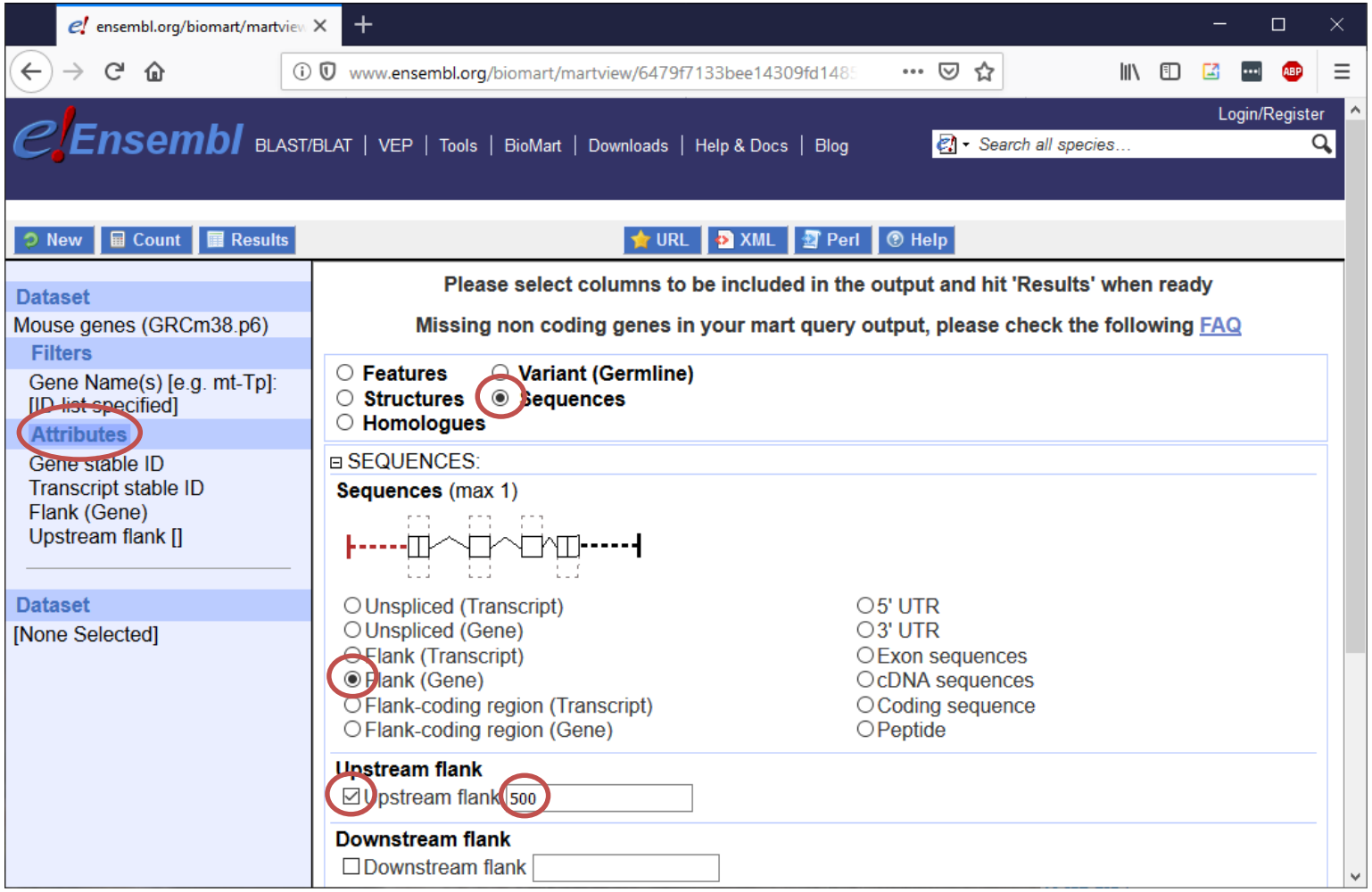


https://ensembl.org/biomart/martview

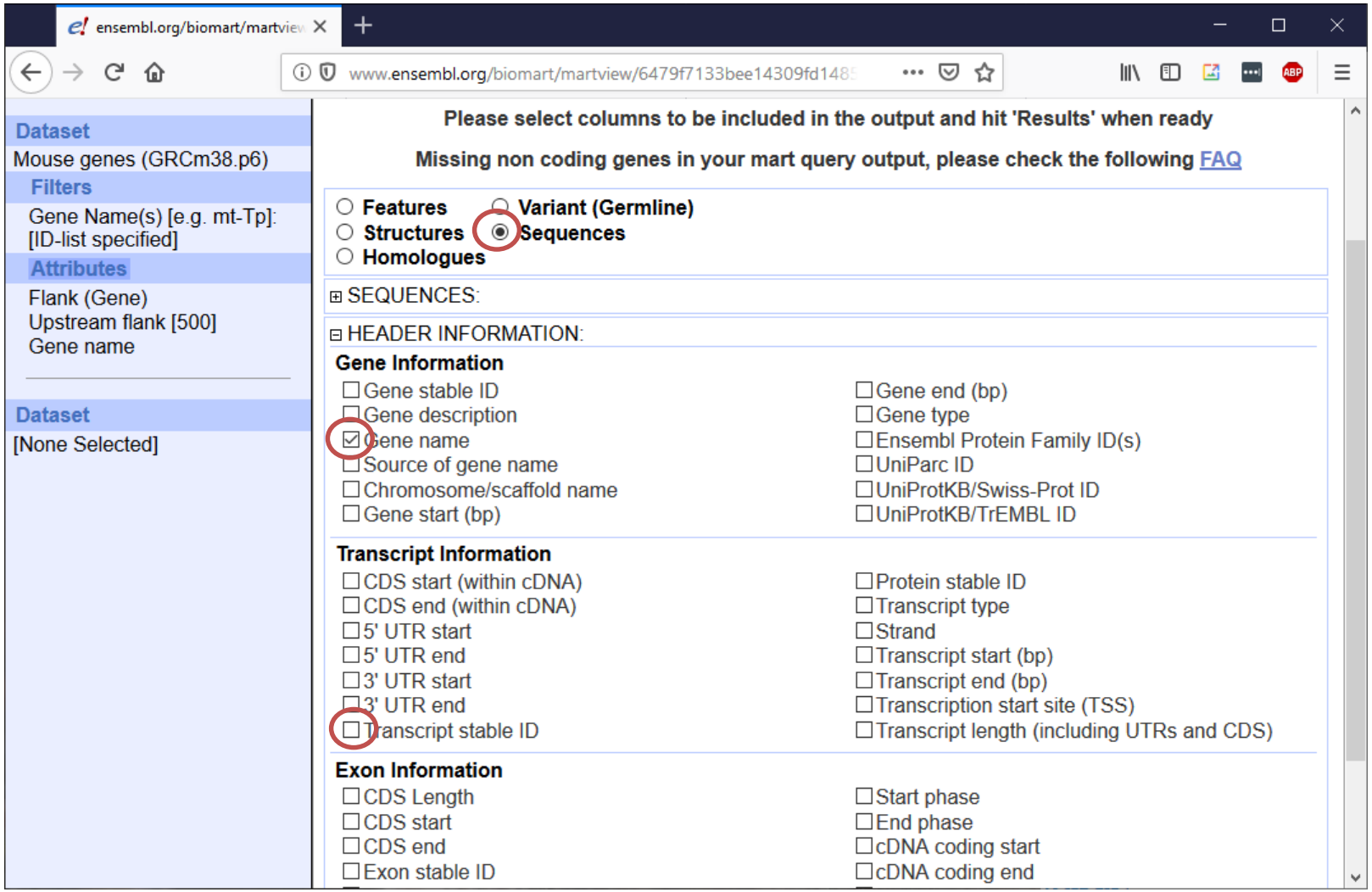
BioMart – Specifying features



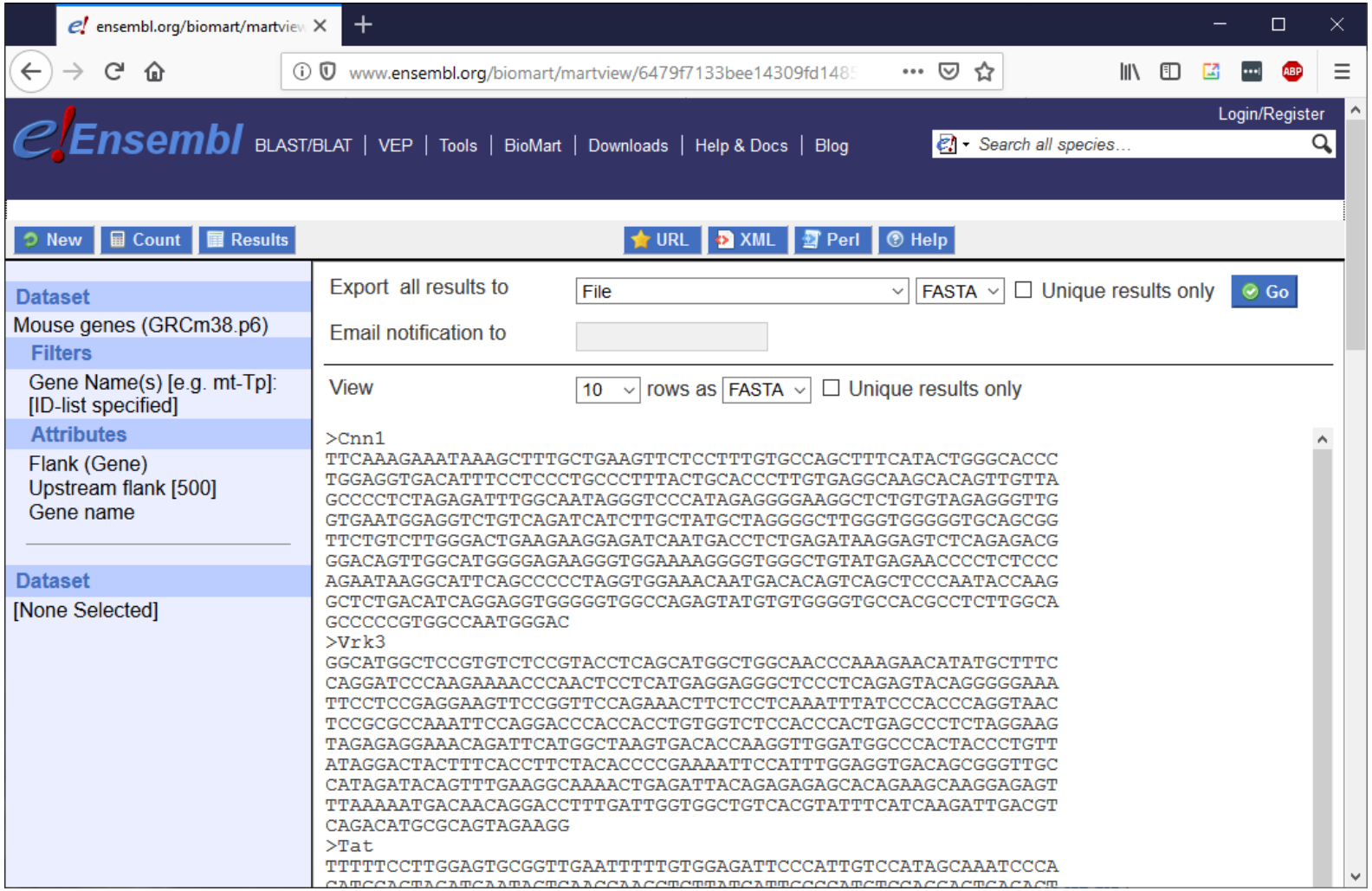
BioMart – selecting seq region



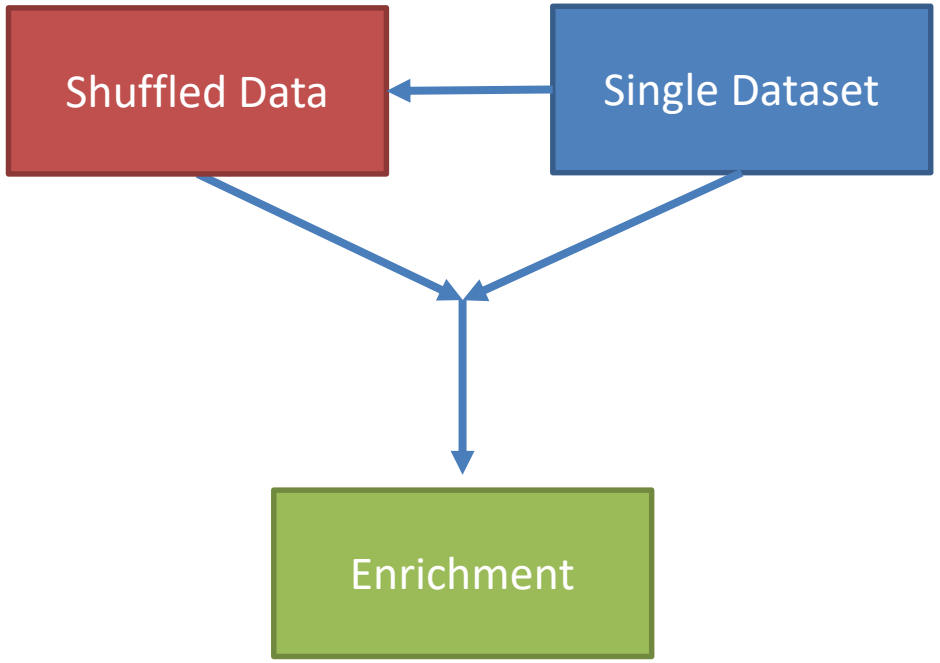
BioMart – header info



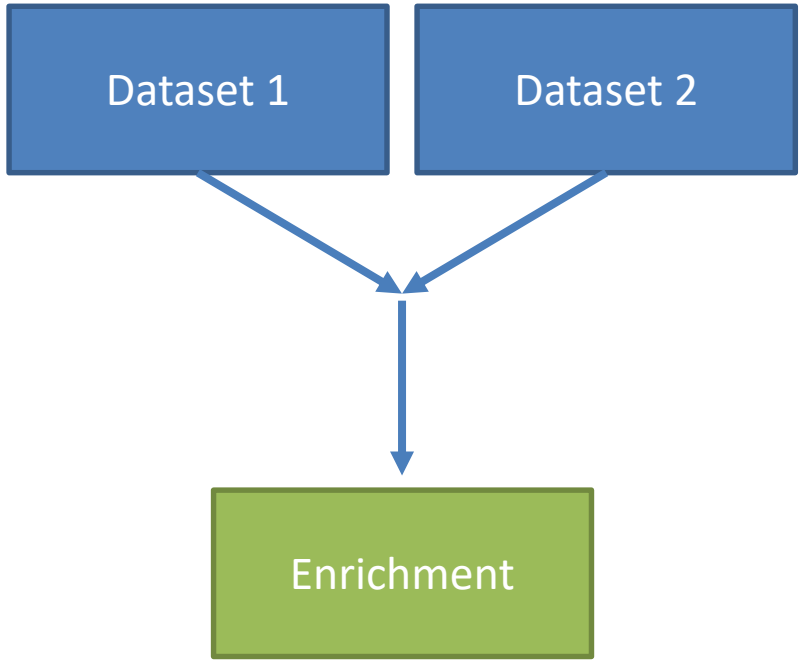
BioMart - exporting



Deciding on a comparison



Single Input Set



Double Input Set

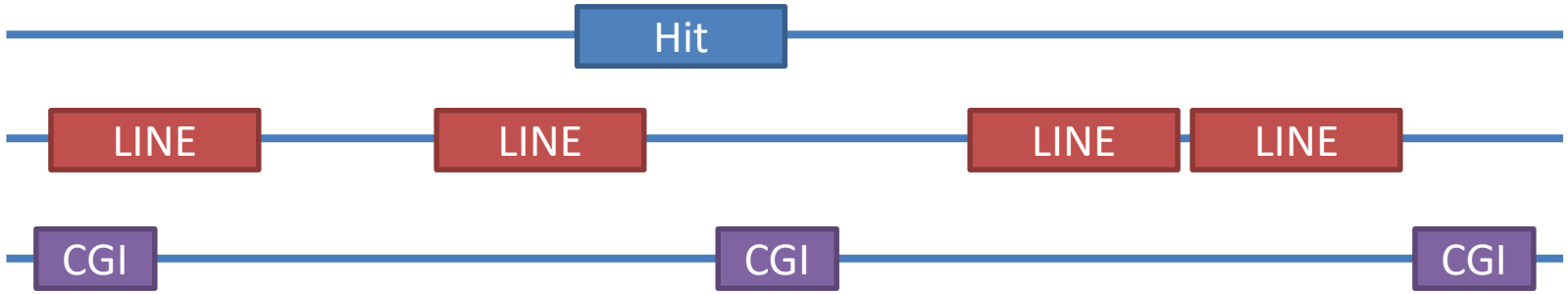
Filtering list of hits

- High specificity
- Quick run times
- Potentially lower power
- Highest hit artefacts

- More power
- Long run times
- More noise

- Don’t need all hits to generate motifs
- Often better to have a smaller, cleaner sequence set

Artefacts



- Exclude common repeats
  - Simple repeats (poly-A, SerThr repeats etc)
  - Complex repeats (retroviral etc)

RepeatMasker

- Check composition
  - Analyse compositionally biased regions explicitly

Compter
Nucleotide k-mer clustering

Software

The MEME Suite
Motif-based sequence analysis tools

meme-suite.org

XXmotif

xxmotif.genzentrum.lmu.de/

CisFinder

lgsun.grc.nia.nih.gov/CisFinder/

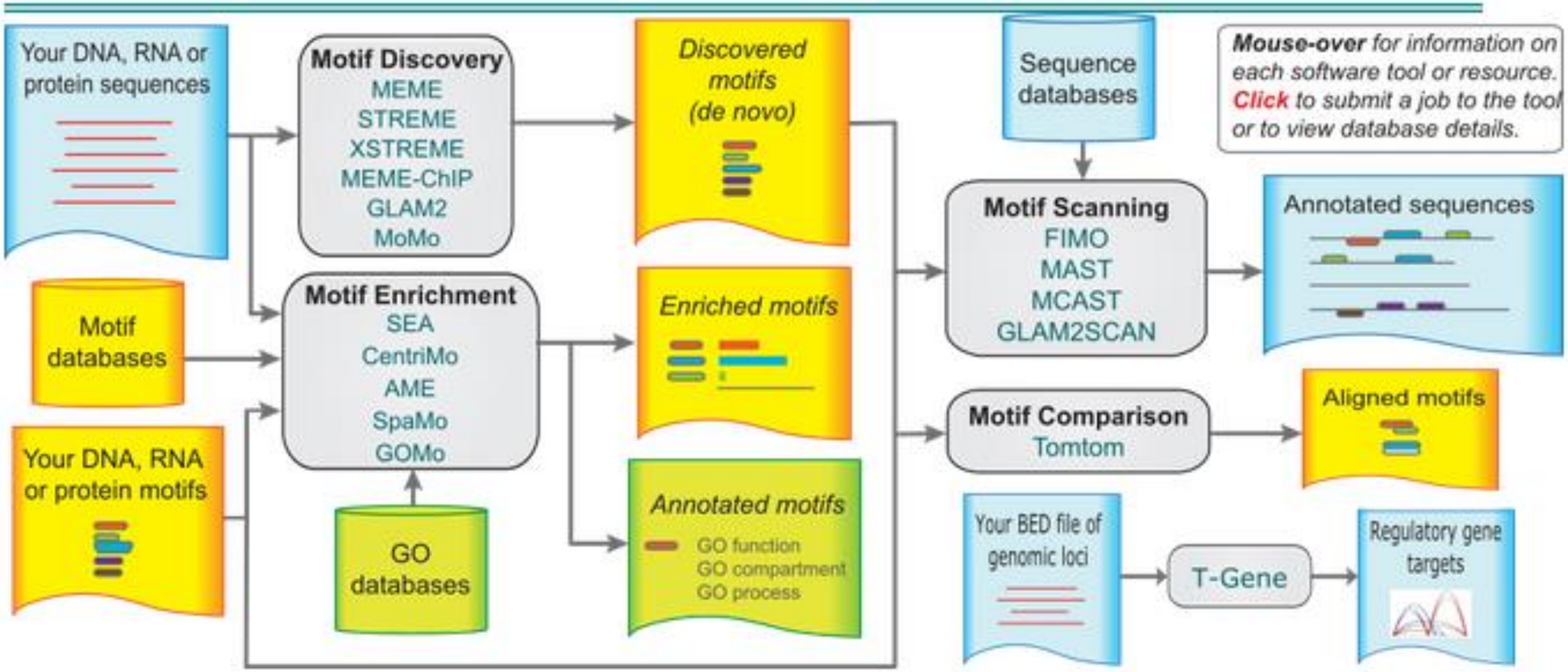
CREAD

cb.utdallas.edu/cread/

HOMER

homer.salk.edu/homer/motif/

MEME Suite

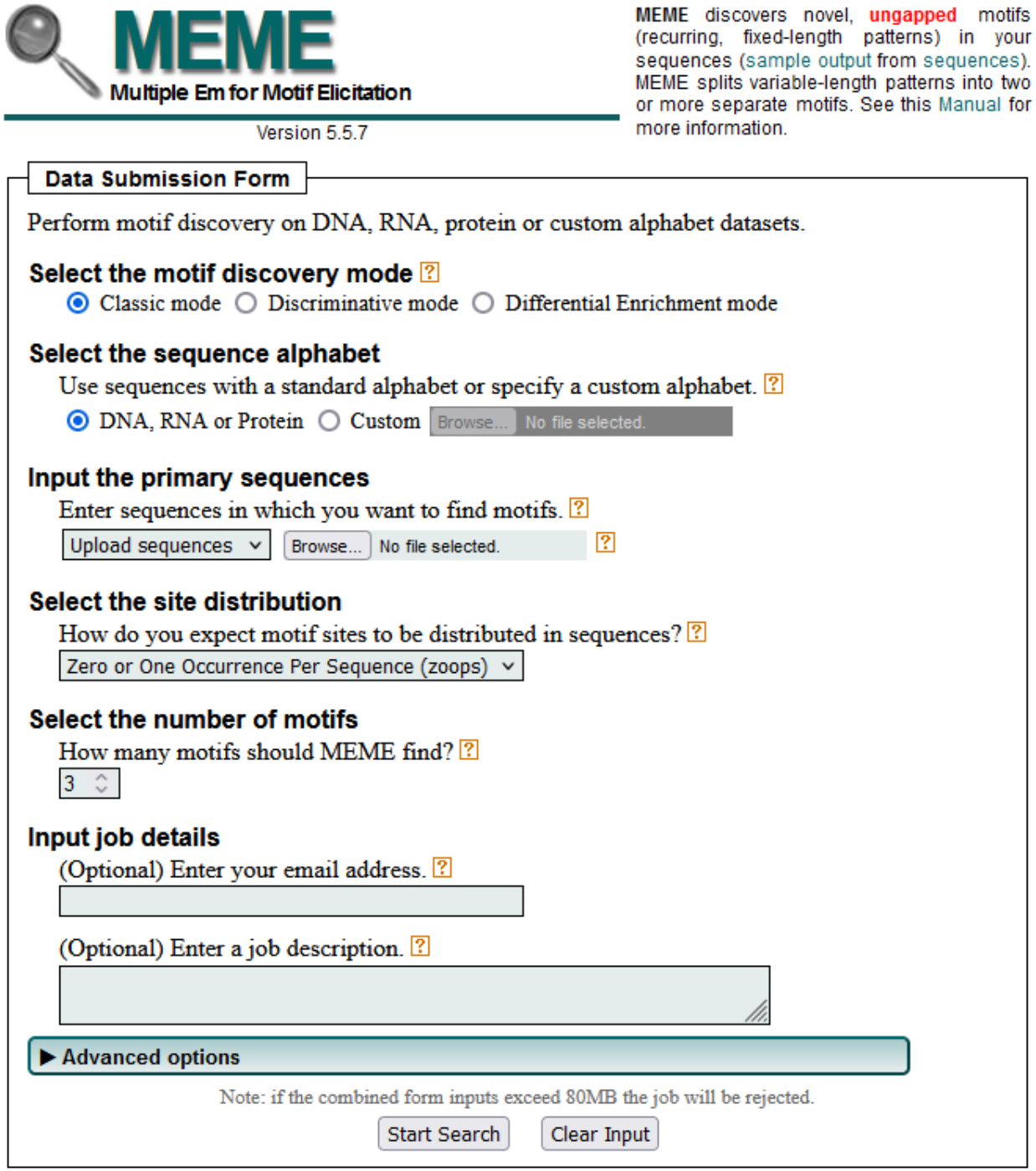

Your DNA, RNA or protein sequences
Motif Discovery
MEME
STREME
XSTREME
MEME-ChIP
GLAM2
MoMo
Discovered motifs (de novo)
Sequence databases
Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.
Motif databases
Motif Enrichment
SEA
CentriMo
AME
SpaMo
GOMo
Enriched motifs
Motif Scanning
FIMO
MAST
MCAST
GLAM2SCAN
Annotated sequences
Your DNA, RNA or protein motifs
GO databases
Annotated motifs
GO function
GO compartment
GO process
Motif Comparison
Tomtom
Aligned motifs
Your BED file of genomic loci
T-Gene
Regulatory gene targets

MEME Motif Discovery

- MEME
  - Original motif enrichment program
  - PWM based motifs
  - Long ungapped motifs, sensitive search, slow!

- STREME/XSTREME
  - Short ungapped discriminatory motifs
    - STREME when you expect the motif to be positioned within your sequence (ie ChIP peaks)
    - XSTREME when you don't expect the motif to be positioned (eg Promoters)
  - Degeneracy based motifs
  - Quick!

- GLAM2
  - Gapped motifs



Main Parameters:

- Sequences (multi-fasta)
- Expected sites
- How many motifs to find

Advanced

- Custom background
- Negative set
- Motif size restriction

NB: Query size limited to 60kb

Local installations don’t have this limit

Good Result

MEME

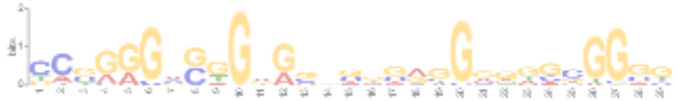
Multiple Em for Motif Elicitation

For further information on how to interpret these results or to get a copy of the MEME software please access http://meme-suite.org.

If you use MEME in your research, please cite the following paper:
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994. [pdf]

DISCOVERED MOTIFS | MOTIF LOCATIONS | PROGRAM INFORMATION

DISCOVERED MOTIFS

	Logo	E-value	Sites	Width	More	Submit/Download
1.		2.2e-009	25	11		
2.		7.5e+001	25	15		
3.		2.3e+003	13	11		
4.		2.4e+004	18	8		
5.		4.0e+005	23	29		

Good Result - Motif

Discovered Motifs

1.

***E*-value:** 2.2e-009 **Site Count:** 25 **Width:** 11

Standard | Reverse Complement

Log Likelihood Ratio: 259 **Information Content:** 15.7 **Relative Entropy:** 14.9 **Bayes Threshold:** 9.99832

	Name	Strand	Start	*p*-value	Sites
10.	Chr1:36554101-36554850	+	377	3.43e-7	TGAGGGCGGC ACCGGAAGTGG CGAGCAGTCT
9.	Chr1:33592501-33593400	+	695	3.43e-7	CCAGGGTAAC ACCGGAAGTGG GCTTATTTGC
24.	Chr1:151881751-151882500	-	438	2.41e-6	TTTGAGAAGC CCCGGAAGTGG CCCGGCTGTT
17.	Chr1:46153351-46154250	+	507	2.41e-6	GAATCCACTT GCCGGAAGTGC CTTTCCAGTG
11.	Chr1:38455351-38456250	+	405	2.41e-6	TTCCGCGAGT GCCGGAAGTGC CCCGGGACTT
5.	Chr1:20987551-20988450	+	459	2.41e-6	GGAGAGCGCC GCCGGAAGTGC CCTGGTGGGA
3.	Chr1:10003051-10003950	-	411	2.41e-6	GTTCCAGGGG GCCGGAAGTGC GTGCGTCCAG
8.	Chr1:28969201-28969950	-	396	2.99e-6	GGGGAGCGGT GCCGGAAGTAG TCGGGGAGGG
4.	Chr1:20834251-20835150	+	421	3.83e-6	CGACGCCGGA ACCGGAAGCGG GGCAAGGGGT

Good Result - Positioning

Motif Locations

Only Motif Sites Motif Sites/Scan All Sequences

	Name	p-value	Motif Location
1.	Chr1:1342201-1343100	1.73e-5	
2.	Chr1:2322751-2323650	6.33e-5	
3.	Chr1:10003051-10003950	1.10e-6	
4.	Chr1:20834251-20835150	7.60e-6	
5.	Chr1:20987551-20988450	8.59e-10	
6.	Chr1:27648301-27649200	1.52e-2	
7.	Chr1:28655101-28656000	2.85e-4	
8.	Chr1:28969201-28969950	2.36e-5	
9.	Chr1:33592501-33593400	1.69e-8	
10.	Chr1:36554101-36554850	9.04e-7	
11.	Chr1:38455351-38456250	3.12e-8	

TTCCGCGAGT GCCGGAAGTGC CCCGGGACTT

Motif 1

***p*-value** 2.41e-6

Start 405

End 415

MEME version

4.10.1 (Release date: Wed Mar 25 11:40:43 2015 +1000)

Reference

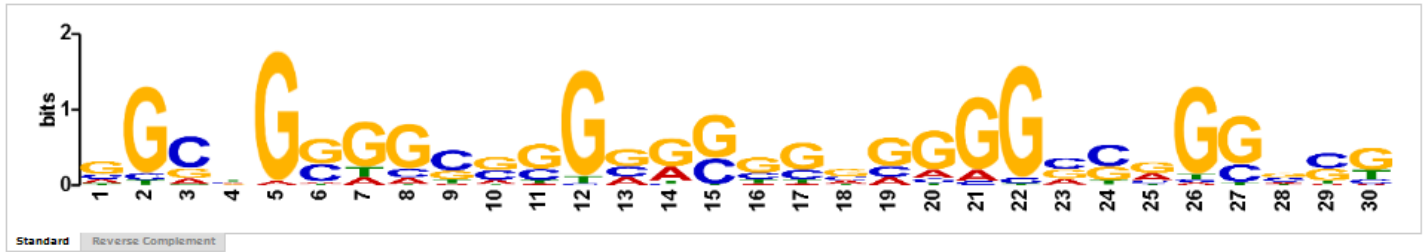
Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second*

For ‘peak’ data, expect motifs to be roughly centred. For promoter data there may be no pattern.

Artefactual Result - Composition

Discovered Motifs

1.
E-value: 2.6e-035 Site Count: 63 Width: 30

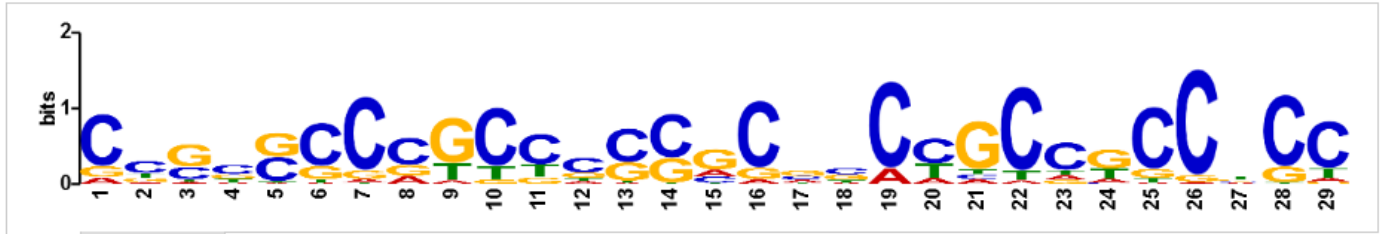


Standard Reverse Complement

Log Likelihood Ratio: 790 Information Content: 21.3 Relative Entropy: 18.1 Bayes Threshold: 11.037

Name	Strand	Start	*p*-value	Sites
27. Chr22:42720298-42721498	+	788	2.11e-11	GCCTCGGGAA GGCAGGGGCAGGGGCGGAGGGGCTGGGCCG CCGCCCTAGC
67. Chr11:62760544-62761744	+	879	3.68e-10	GTCGTTTGAT GGCTGGGGGGCGGGGGTGAGGGGCGGGGCT TAGTGGTTCT
26. Chr17:29929276-29930476	+	183	1.58e-9	GTTACGACCA GGCGGCGGCGCGCGCCGAGGGGAGAGGCGG TAGCTGACAG
73. Chr16:667086-668286	+	980	3.34e-9	GCGGCTGTGC GGGCGGGGCCGGAAGTGTGGGGACAGCGCG GGGCCGAGAA
53. Chr19:16323817-16325017	-	1017	5.14e-9	CCCGAGGCCG GGCTGGTCGGGGCGCGGGCGGGCGGGGACG GCTCTGTGGG
32. Chr21:41952890-41954090	+	96	5.91e-9	GCCACCGCTG GGCCGCGGGCGGGACGCCCGGGGCGGCGCG CGCAGGAGGA
43. Chr18:79963561-79964761	-	1007	1.33e-8	TCCAGGGCGG AGGCGGGGCCGGAGCGGAGAGGCGGGCGCT GCACCTGCAG
12. Chr3:47781106-47782306	+	515	1.72e-8	CGACGATGGC CGCAGCGGCGGGCGGCGGCGGGCCGGGGAC AGCGGTAGGC

2.
E-value: 2.0e-021 Site Count: 51 Width: 29



Standard Reverse Complement

Log Likelihood Ratio: 660 Information Content: 21.7 Relative Entropy: 18.7 Bayes Threshold: 11.1181

Name	Strand	Start	*p*-value	Sites
38. Chr12:42589355-42590555	+	834	9.20e-11	GCCCAGCTGC CAGCGCCGTCTCCCGCGGCCGCCGCCGCC CGACAGCGCG
26. Chr17:29929276-29930476	+	332	8.41e-10	GGTGCGGAGT CTGCGGCCGTTCCCGCGGCCTCCTCCTCC NNNNNNNNNN
32. Chr21:41952890-41954090	-	746	1.50e-9	CCTCGGCCCG GACCGCCCGCCCCGGCATCTGCCGCCCCC CGGCCCGCAG
49. Chr19:19319681-19320881	+	1115	2.27e-9	GAGGAGAGCC CGGCCCCGTCTCGCGCGCCTGCTTCCGCC TCCCTGTGGC
56. Chr9:137187647-137188847	+	856	7.47e-9	GAGTCGCTGT CCCCCTCCGCCACCACAACTGCCGCCGCC ATCTGCACCC
4. Chr1:23176808-23178008	+	280	1.23e-8	CAAACGGCGA ACGCGCCGGCCCCCTCCCCCGCCGCCCGC AAGCCTGGTG
21. Chr22:38397522-38398722	+	307	1.57e-8	CGCCCCCCGC CCGCGCCAGCCGGGGCGCCCGCCTGGTCC TGCGGAGGCT
27. Chr22:42720298-42721498	+	874	2.00e-8	CACGAACCCC GCGCGGCGGTCCCGGCCACCGCCTCCACC CCAACTGCGC

MEME tends to favour long compositionally biased motifs
Real motifs can be further down the list

Artefactual Result - Duplication

3.

E-value: 4.8e-022 Site Count: 5 Width: 49

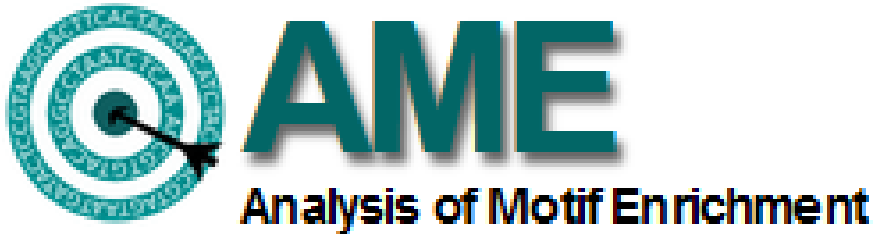
Log Likelihood Ratio: 266 Information Content: 76.9 Relative Entropy: 76.9 Bayes Threshold: 11.863

	Name	Strand	Start	*p*-value	Sites
24.	Tbc1d14	+	57	3.43e-30	TTGGTTTTTG AGACAGGGTTTCTCTGTGTAGCCCTGGCTGTCCTGGAACTCACTCTGTA GACCAGGCTG
5.	Cd79a	-	18	9.27e-30	TTGTTTTTCG AGACAGGGTTTCTCTGTGTAGCCCTGGCTGTCCTGGAACTCATTCTGTA GACCAGGCTG
61.	Cpne1	+	96	1.79e-28	TGGTTTTTCG AGACAGGGTTTCTCTGTATAGTCCTGGCTGTCCTGGAACTCACTCTGTA GCCCAGGCTG
17.	H3f3b	-	128	8.15e-24	TTTTTTTTTG AGACAGGGTTTCTCTGTATAGCCCTGGCTGTCCTGGAACGACCAGGCTG GCCTCGAACT
39.	Zfp410	+	6	6.93e-20	TTTTG GACAAGGTCTCAACTATGTAGCCCTCGTTCACCGGGAACTCCTGGAGTA CTGGGATTGC

Multiple transcripts with the same promoter

Overlapping regions

AME – Known motif search



- Quicker / easier than de-novo discovery
- Limited to characterised binding sites
- Can choose from common motif sources
- Good place to start

AME

Analysis of Motif Enrichment

Data Submission Form

Perform standard (non-local) motif enrichment analysis.

Select the type of control sequences to use

(•) Shuffled input sequences ( ) User-provided control sequences ( ) NONE

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet.

(•) DNA, RNA or Protein ( ) Custom [Browse...] No file selected.

Input the primary sequences

Enter the sequences in which you want to find enriched motifs.

[Upload sequences] [Browse...] No file selected.

Input the motifs

Select a motif database or enter the motifs you wish to test for enrichment.

[Eukaryote DNA] DNA

[Vertebrates (In vivo and in silico)]

Select the sequence scoring method

[Average odds score]

Select the motif enrichment test

[Fisher's exact test]

Input job details

(Optional) Enter your email address.

(Optional) Enter a job description.

▶ Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

[Start Search] [Clear Input]

Databases (select category)

- Eukaryote DNA
- Prokaryote DNA
- Methylcytosine DNA
- JASPAR (NON-REDUNDANT) DNA
- JASPAR (REDUNDANT) DNA
- JASPAR COLLECTIONS DNA
- HOCOMOCO (HUMAN + MOUSE orthologs) DNA
- TFBSshape DNA
- CIS-BP 2.00 Single Species DNA
- CIS-BP 1.02 Single Species DNA
- ARABIDOPSIS (Arabidopsis thaliana) DNA
- ECOLI (Escherichia coli) DNA
- FLY (Drosophila melanogaster) DNA

Databases

- JASPAR CORE (2022)
- JASPAR CORE (2022) vertebrates
- JASPAR CORE (2022) fungi
- JASPAR CORE (2022) insects
- JASPAR CORE (2022) nematodes
- JASPAR CORE (2022) plants
- JASPAR CORE (2022) urochordates

AME Result

No additional detail

Could check for positional bias with CentriMo

Beware similar motifs from different factors

AME
Analysis of Motif Enrichment

For further information on how to interpret these results or to get a copy of the MEME software please access http://meme-suite.org.

If you use AME in your research, please cite the following paper:
Robert McLeay and Timothy L. Bailey, "Motif Enrichment Analysis: A unified framework and method evaluation", *BMC Bioinformatics*, **11**:165, 2010, doi:10.1186/1471-2105-11-165.
[full text]

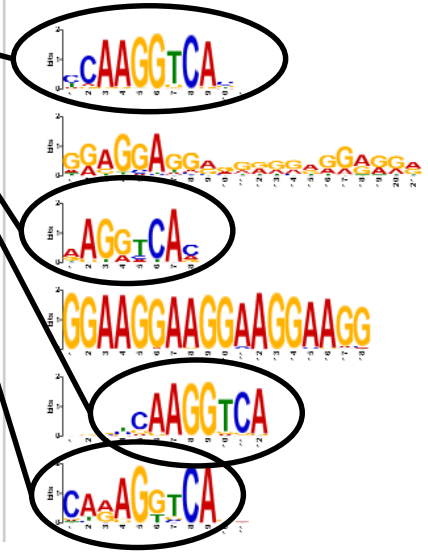
ENRICHED MOTIFS | INPUT FILES | PROGRAM INFORMATION

ENRICHED MOTIFS

Fixed partition size: number of primary sequences (99)

Sequence motif score: avg_odds
Background model source file: motif input file
Background model frequencies: 0.25,0.25,0.25,0.25
Total pseudocount added to a motif column: 0.25

Statistical test: Wilcoxon rank-sum test
Ranksum method: quick
Threshold *p*-value for reporting results: 0.05
Number of multiple tests for Bonferroni correction: #Motifs × #PartitionsTested = 205 × 1 = 205



Logo	Database	ID	Name	*p*-value	Adjusted *p*-value
	JASPAR CORE 2014 vertebrates	MA0592.1	ESRRA	5.49e-10	1.13e-7
	JASPAR CORE 2014 vertebrates	MA0528.1	ZNF263	5.26e-7	1.08e-4
	JASPAR CORE 2014 vertebrates	MA0160.1	NR4A2	2.73e-6	5.59e-4
	JASPAR CORE 2014 vertebrates	MA0149.1	EWSR1-FLI1	3.37e-6	6.90e-4
	JASPAR CORE 2014 vertebrates	MA0141.2	Esrrb	4.99e-6	1.02e-3
	JASPAR CORE 2014 vertebrates	MA0512.1	Rxra	9.82e-6	2.01e-3

MEME Suite 5.4.1

▼Motif Discovery
- MEME
- STREME
- XSTREME
- MEME-ChIP
- GLAM2
- MoMo
- DREME (deprecated)

►Motif Enrichment
►Motif Scanning
►Motif Comparison
►Gene Regulation
►Manual
►Guides & Tutorials
►Sample Outputs
►File Format Reference
►Databases
►Download & Install
►Help
►Alternate Servers
►Authors & Citing
►Recent Jobs

↳ Previous version 5.3.3

STREME

Sensitive, Thorough, Rapid, Enriched Motif Elicitation

Version 5.4.1

STREME discovers ungapped motifs (recurring, fixed-length patterns) that are enriched in your sequences or relatively enriched in them compared to your control sequences (sample output from sequences). See this Manual or this Tutorial for more information.

Data Submission Form

Perform discriminative motif discovery in sequence datasets (including in very **large** datasets). The sequences may be in the DNA, RNA or protein alphabet, or in a custom alphabet.

Select the type of control sequences to use

○ Shuffled input sequences ◉ User-provided sequences ?

Select the sequence alphabet

Use sequences with a standard alphabet or specify a custom alphabet. ?

◉ DNA, RNA or Protein ○ Custom Browse... No file selected.

Input the sequences

Enter the sequences in which you want to find motifs. ?

Upload sequences Browse... positive_cgi_set.txt DNA ?

Input the control sequences

STREME will find motifs that are enriched relative to these sequences. ?

Upload sequences Browse... negative_cgi_set.txt DNA ?

Convert DNA sequences to RNA?

☐ Convert DNA to RNA ?

Input job details

(Optional) Enter your email address. ?

(Optional) Enter a job description. ?

► Advanced options

Note: if the combined form inputs exceed 80MB the job will be rejected.

Start Search Clear Input

STREME Results

https://meme-suite.org/meme//opal-jobs/appSTREME_5.4.1163733537444312168004 90%

STREME

Sensitive, Thorough, Rapid, Enriched Motif Elicitation

For further information on how to interpret these results please access https://meme-suite.org/meme/doc/streme.html.

To get a copy of the MEME software please access https://meme-suite.org.

Details

Train Positives	Train Negatives	Score	Test Positives
69 / 81 (85.2%)	2 / 83 (2.4%)	7.2e-031	6 / 8 (75.0

3-CCCTGGGGCGS			1.0e+000	4.0e+0

Submit or Download

Submit Motif | Download Motif | Download Logo

Submit to program

◉	Tomtom	Find similar motifs in published libraries or a library you supply.
○	FIMO	Find motif occurrences in sequence data.
○	MAST	Rank sequences by affinity to groups of motifs.
○	GOMo	Identify possible roles (Gene Ontology terms) for motifs.
○	SpaMo	Find other motifs that are enriched at specific close spacings which might imply the existence of a complex.

Submit Cancel

Tomtom

Motif Comparison Tool

For further information on how to interpret these results or to get a copy of the MEME software please access http://meme-suite.org.

If you use TOMTOM in your research, please cite the following paper:
Shobhit Gupta, JA Stamatoyannopolous, Timothy Bailey and William Stafford Noble, "Quantifying similarity between motifs", *Genome Biology*, 8(2):R24, 2007. [full text]

QUERY MOTIFS | TARGET DATABASES | MATCHES | PROGRAM INFORMATION

QUERY MOTIFS

Next Top

Name	Alt. Name	Preview	Matches	List
GCCTCTAA	DREME		3	MA0503.1 (Nkx2-5), MA0122.1 (Nkx3-2), MA0504.1 (NR2C2)

TARGET DATABASES

Previous Next Top

Database	Number of Motifs	Motifs Matched
JASPAR_CORE_2014_vertebrates.meme	205	3

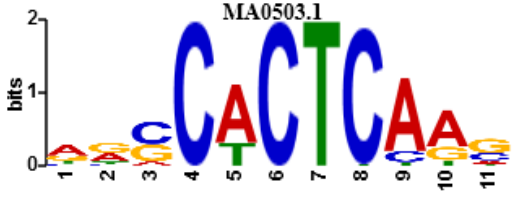
MATCHES TO QUERY MOTIF GCCTCTAA (DREME)

Previous Next Top

Summary

Name	MA0503.1
Alt. Name	Nkx2-5
Database	JASPAR_CORE_2014_vertebrates.meme
***p*-value**	0.0152266
***E*-value**	3.12146
***q*-value**	1
Overlap	8
Offset	1
Orientation	Normal

Alignment



GCCTCTAA

Motif Searching Exercise

Babraham Bioinformatics