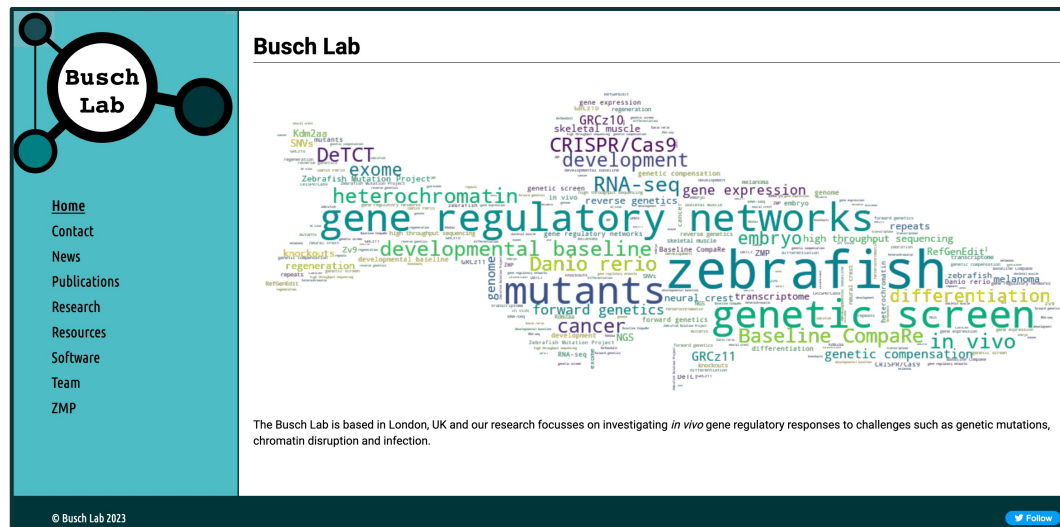


# Introduction to RNA-seq and functional interpretation: Next steps in gene prioritisation

22nd Feb 2023

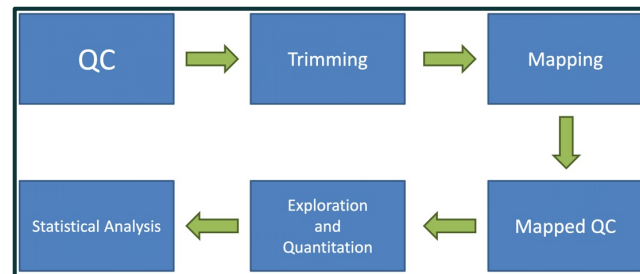


- Ian Sealy
- Busch Lab, QMUL
- Previously at Sanger Institute
- RNA-seq / zebrafish
- Run “*Bioinformatics & Functional Genomics in Zebrafish*” course at EBI

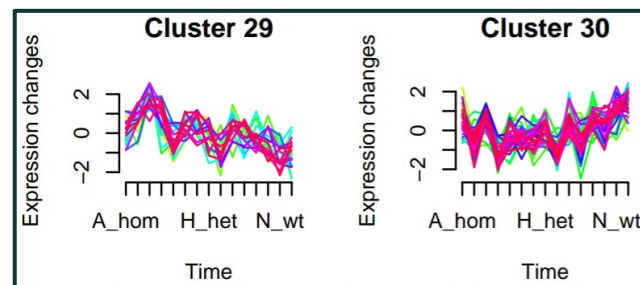


# Gene list of interest

- Starting point for today: **gene list of interest**
- Most likely from RNA-seq differential expression analysis
- But could be a list from any other analysis:
  - Clustering genes with similar expression profiles
  - Microarray analysis
  - Quantitative proteomics
  - Differential methylation analysis
  - etc...



From: Simon Andrews (yesterday)



# Unranked or ranked gene list?

- Gene list can be:
  - Unranked (e.g. genes with somatic mutations in cancer sample)
  - Ranked (e.g. sensitivity in a CRISPR screen)
- RNA-seq differential expression analysis produces ranked lists
- Ranked lists are ordered by a score or metric:
  - e.g. adjusted p-value
  - e.g.  $\log_2$  fold change
- Ranked lists can also have a threshold applied:
  - e.g. adjusted p-value < 0.05

```
ENSDARG00000043198
ENSDARG00000075229
ENSDARG00000036695
ENSDARG00000092115
ENSDARG00000013076
ENSDARG00000015890
ENSDARG00000060682
ENSDARG00000076241
ENSDARG00000093347
ENSDARG00000098114
```

```
ENSDARG00000075676 0.039
ENSDARG00000104197 0.041
ENSDARG00000004301 0.041
ENSDARG00000079766 0.042
ENSDARG00000030494 0.042
ENSDARG00000116804 0.043
ENSDARG00000100599 0.043
ENSDARG00000104325 0.043
ENSDARG00000111102 0.043
ENSDARG00000022466 0.044
```



# “Gene” list of interest

- May not actually be a list of genes
- Could be transcripts or proteins or SNPs, etc...
- Most tools require a list of genes so need to convert
- BioMart is a useful tool for conversions (and other bioinformatics tasks):  
[www.ensembl.org/biomart/martview](http://www.ensembl.org/biomart/martview)

New Count Results URL XML Perl Help

**Dataset 5 / 69299 Genes**  
Human genes (GRCh38.p13)

**Filters**  
Transcript stable ID(s) [e.g. ENST00000000233]: [ID-list specified]

**Attributes**  
Gene stable ID  
Gene name  
Gene description

**Dataset**  
[None Selected]

**Please restrict your query using criteria below**  
(If filter values are truncated in any lists, hover over the list item to see the full text)

**REGION:**

**GENE:**

☐ Limit to genes (external references)... ☒ Only ☐ Excluded

☒ Input external references ID list [Max 500 advised]

With BioGRID Interaction data, The General Repository

ENST00000410016  
ENST00000683222  
ENST00000593259  
ENST00000444827  
ENST00000577092

# What next?

- Have a gene list, but what do you do next?
- How do you relate the gene list to existing knowledge?

Gene	pval	adjp	log2fc
ENSDARG00000041294	4.904002310063973e-37	1.0867269119101765e-32	1.5709251030700861
ENSDARG00000060498	1.1297090308658515e-25	1.2517176061993635e-21	1.5921762041345
ENSDARG00000031683	3.2009883731403506e-25	2.364463411626339e-21	-1.277820860357806
ENSDARG00000077982	5.3336179195843655e-18	2.9548243274497384e-14	0.9349522690823255
ENSDARG00000070480	1.2940060161760502e-17	5.735034663692255e-14	1.0699010828953783
ENSDARG00000007769	4.245003753873642e-17	1.5678213864306653e-13	1.6785196633873156
ENSDARG00000102435	6.025610180317608e-17	1.9075360227976884e-13	1.0539265022132713
ENSDARG00000101482	9.742460938723084e-17	2.6986616800262944e-13	0.9350743176658163
ENSDARG00000034503	2.261103100242347e-16	5.567338300152267e-13	0.6082489350504545

# What next?


- Have a gene list, but what do you do next?
- How do you relate the gene list to existing knowledge?
- Add annotation (e.g. BioMart)

Gene	pval	adjp	log2fc
ENSDARG00000041294	4.904002310063973e-37	1.0867269119101765e-32	1.5709251030700861

Gene	pval	adjp	log2fc	Chr	Start	End	Name	Description
ENSDARG00000041294	4.904002310063973e-37	1.0867269119101765e-32	1.5709251030700861	3	62161184	62169060	noxo1a	NADPH oxidase organizer 1a
ENSDARG00000060498	1.1297090308658515e-25	1.2517176061993635e-21	1.5921762041345	23	30006206	30010042	tnfrsf9a	tumor necrosis factor receptor superfamily, member 9a
ENSDARG00000031683	3.2009883731403506e-25	2.364463411626339e-21	-1.277820860357806	20	46552311	46554440	fosab	v-fos FBJ murine osteosarcoma viral oncogene homolog Ab
ENSDARG00000077982	5.3336179195843655e-18	2.9548243274497384e-14	0.9349522690823255	22	661505	665371	elf3	E74-like factor 3 (ets domain transcription factor, epithelial-specific)
ENSDARG00000070480	1.2940060161760502e-17	5.735034663692255e-14	1.0699010828953783	19	30400372	30404096	agr2	anterior gradient 2
ENSDARG00000007769	4.245003753873642e-17	1.5678213864306653e-13	1.6785196633873156	7	56602521	56606752	sult5a1	sulfotransferase family 5A, member 1
ENSDARG00000102435	6.025610180317608e-17	1.9075360227976884e-13	1.0539265022132713	7	45975537	45976956	plekhf1	pleckstrin homology domain containing, family F (with FYVE domain) member 1
ENSDARG00000101482	9.742460938723084e-17	2.6986616800262944e-13	0.9350743176658163	5	13870340	14004206	hk2	hexokinase 2
ENSDARG00000034503	2.261103100242347e-16	5.567338300152267e-13	0.6082489350504545	2	48309600	48375342	per2	period circadian clock 2

ENSDARG00000102435	6.025610180317608e-17	1.9075360227976884e-13	1.0539265022132713
ENSDARG00000101482	9.742460938723084e-17	2.6986616800262944e-13	0.9350743176658163
ENSDARG00000034503	2.261103100242347e-16	5.567338300152267e-13	0.6082489350504545

# Look up genes in databases

GENE	
<b><i>noxo1a</i></b>	
ID	ZDB-GENE-030131-9700
Name	<i>NADPH oxidase organizer 1a</i>
Symbol	<i>noxo1a</i> <a href="#">Nomenclature History</a>
Previous Names	<i>noxo1</i> , <i>cb18</i> (1), <i>sb:cb18</i> , <i>SNX28b</i> (1), <i>wu:fd09d09</i> , <i>zgc:152911</i> (1)
Type	<a href="#">protein_coding_gene</a> <a href="#">↗</a>
Location	Chr: 3 <a href="#">Mapping Details/Browsers</a>
Description ⓘ	Predicted to have phosphatidylinositol-3-phosphate binding activity and superoxide-generating NADPH oxidase activator activity. Predicted to be involved in superoxide metabolic process. Predicted to localize to NADPH oxidase complex and cytoplasm. Is expressed in EVL; periderm; and pharynx. Orthologous to human <a href="#">NOXO1</a> (NADPH oxidase organizer 1).
Genome Resources	<a href="#">Alliance</a> <a href="#">↗</a> (1), <a href="#">Gene:572245</a> <a href="#">↗</a> (1), <a href="#">Ensembl(GRCz11):ENSDARG00000041294</a> <a href="#">↗</a> (3)
Note	None
Comparative Information	

# Look up genes in databases

GENE	
noxo	● GENE
ID	NOXO1
Name	
Symbol	
Previous N	
Type	
Location	
Description	
Genome Re	
Note	
Comparativ	
Information	
Species	<i>Homo sapiens</i>
Symbol	NOXO1
Name	NADPH oxidase organizer 1
Synonyms	MGC20258 NADPH oxidase regulatory protein <a href="#">▼ Show All 12</a>
Biotype	protein coding gene
Automated Description ?	Enables enzyme binding activity. Involved in extracellular matrix disassembly. Part of NADPH oxidase complex.
RGD Description	This gene encodes an NADPH oxidase (NOX) organizer, which positively regulates NOX1 and NOX3. The protein contains a PX domain and two SH3 domains. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene. [provided by RefSeq, Jun 2012]
Cross References	<a href="#">ENSEMBL:ENSG00000196408</a> <a href="#">↗</a> <a href="#">NCBI_Gene:124056</a> <a href="#">↗</a> <a href="#">▼ Show All 4</a>
Additional Information	<a href="#">Literature</a> <a href="#">↗</a>

# Look up genes in databases

GENE

noxo

ID

Name

Symbol

Previous N

Type

Location

Description

Genome Re

Note

Comparativ

Information

RGD Descrip

Cross Refer

Additional Information

GENE

NOXO1

Species

Symbol

Name

Synonyms

Biotype

Automated

RGD Descrip

Cross Refer

Additional Information

Summaries for NOXO1 Gene

?

Entrez Gene Summary for NOXO1 Gene

This gene encodes an NADPH oxidase (NOX) organizer, which positively regulates NOX1 and NOX3. The protein contains a PX domain and two SH3 domains. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene. [provided by RefSeq, Jun 2012]

GeneCards Summary for NOXO1 Gene

NOXO1 (NADPH Oxidase Organizer 1) is a Protein Coding gene. Diseases associated with NOXO1 include [Lung Mucoepidermoid Carcinoma](#) and [Phagocyte Bactericidal Dysfunction](#). Among its related pathways are [Signaling by Rho GTPases](#) and [Disease](#). Gene Ontology (GO) annotations related to this gene include *identical protein binding* and *phospholipid binding*. An important paralog of this gene is [SH3PXD2A](#).

UniProtKB/Swiss-Prot Summary for NOXO1 Gene

Constitutively potentiates the superoxide-generating activity of NOX1 and NOX3 and is required for the biogenesis of otoconia/otolith, which are crystalline structures of the inner ear involved in the perception of gravity. Isoform 3 is more potent than isoform 1 in activating NOX3. Together with NOXA1, may also substitute to NCF1/p47phox and NCF2/p67phox in supporting the phagocyte NOX2/gp91phox superoxide-generating activity. ( [NOXO1\\_HUMAN,Q8NFA2](#) )

Gene Wiki entry for NOXO1 Gene

Additional gene information for NOXO1 Gene

[HGNC \(19404\)](#) [NCBI Entrez Gene \(124056\)](#) [Ensembl \(ENSG00000196408\)](#) [OMIM® \(611256\)](#) [UniProtKB/Swiss-Prot \(Q8NFA2\)](#)  
[Open Targets Platform\(ENSG00000196408\)](#)

[Alliance of Genome Resources](#)

Literature

# Look up genes in databases

GENE

*noxo*

ID

Name

Symbol

Previous N

Type

Location

Description

Genome Re

Note

Comparativ

Information

GENE

NOXO1

Species

Symbol

Name

Synonyms

Biotype

Automated

RGD Descrip

Cross Refer

Additional Information

Entrez Gene Summary

GeneCards Summary

UniProtKB/Swiss-Prot

Gene Wiki entry

Additional gene information

Summaries for NOXO1 Gene

Official Symbol

Official Full Name

Primary source

See related

Gene type

RefSeq status

Organism

Lineage

Also known as

Summary

Expression

Orthologs

This gene encodes a protein that contains two SH3 domains. [provided by RefSeq, Jun 2012]

NOXO1 (NADPH oxidase organizer 1) and Phagocyte NADPH oxidase (NOX) gene family annotations related to this gene.

Constitutively active NOXO1 is a protein that is involved in the regulation of the NADPH oxidase complex. Together with NOX1, NOX2, and NOX3, it forms a complex that generates active oxygen species.

NOXO1 (NADPH oxidase organizer 1) is a protein that is involved in the regulation of the NADPH oxidase complex. Together with NOX1, NOX2, and NOX3, it forms a complex that generates active oxygen species.

NOXO1 (NADPH oxidase organizer 1) is a protein that is involved in the regulation of the NADPH oxidase complex. Together with NOX1, NOX2, and NOX3, it forms a complex that generates active oxygen species.

provided by [HGNC](#)

provided by [HGNC](#)

[HGNC:HGNC:19404](#)

[Ensembl:ENSG00000196408](#) [MIM:611256](#); [AllianceGenome:HGNC:19404](#)

protein coding

REVIEWED

[Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

SNX28; P41NOX; P41NOXA; P41NOXB; P41NOXC; SH3PXD5

This gene encodes an NADPH oxidase (NOX) organizer, which positively regulates NOX1 and NOX3. The protein contains a PX domain and two SH3 domains. Alternatively spliced transcript variants encoding multiple isoforms have been observed for this gene. [provided by RefSeq, Jun 2012]

Broad expression in colon (RPKM 2.7), appendix (RPKM 0.9) and 14 other tissues [See more](#)

[mouse](#) [all](#)

# Look up genes in databases

GENE

noxo

GENE

NOXO1

Summaries for NOXO1 Gene

- Manual literature review is OK for a handful of genes
- But what if there are hundreds or thousands?
- We need an automated process

Expression Broad expression in colon (rat Rm 2.7), appendix (rat Rm 0.5) and 14 other tissues [see more](#)

Orthologs [mouse](#) [all](#)

Additional Information





# Functional enrichment analysis

- **Functional enrichment analysis** (or over-representation) systematically relates your data to existing knowledge
- Can help you to:
  - Gain biological insight
  - Generate new hypotheses
  - Validate your experiment


# Functional gene sets

- Existing knowledge is organised into **functional gene sets** in a standardised way, using data from previous experiments
- A functional gene set is a group of genes with a common biological relationship (e.g. annotated to same biological process or involved in same pathway)
- e.g. circadian rhythm:

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	Assigned By	Name
<a href="#">UniProtKB:A0A024QZG3</a>	ATF5	involved_in	<a href="#">GO:0007623</a>    circadian rhythm	<a href="#">ECO:0000265</a>  IEA	<a href="#">GO_REF:0000107</a>	<a href="#">Ensembl</a>	BZIP domain-containing protein
<a href="#">UniProtKB:A0A024QZQ1</a>	SIRT1	involved_in	<a href="#">GO:0007623</a>    circadian rhythm	<a href="#">ECO:0000265</a>  IEA	<a href="#">GO_REF:0000107</a>	<a href="#">Ensembl</a>	Deacetylase sirtuin-type domain-containing protein
<a href="#">UniProtKB:A0A024R230</a>	NTRK2	involved_in	<a href="#">GO:0007623</a>    circadian rhythm	<a href="#">ECO:0000265</a>  IEA	<a href="#">GO_REF:0000107</a>	<a href="#">Ensembl</a>	Tyrosine-protein kinase receptor
<a href="#">UniProtKB:A0A024R241</a>	NFIL3	involved_in	<a href="#">GO:0007623</a>    circadian rhythm	<a href="#">ECO:0000256</a>  IEA	<a href="#">GO_REF:0000002</a>	<a href="#">InterPro</a>	Nuclear factor interleukin-3-regulated protein

# Functional annotation

- Functional annotation is created and maintained by many dedicated databases and projects, e.g.
  - Gene Ontology (GO)
  - Reactome
  - KEGG
  - TRANSFAC



The screenshot shows the Gene Ontology (GO) website. At the top left is the GO logo with the text "GENE ONTOLOGY" and "Unifying Biology". At the top right is a hamburger menu icon. Below the header, a dark blue banner contains the text: "Current release 2023-01-01: 43,248 GO terms | 7,503,460 annotations | 1,475,947 gene products | 5,267 species (see statistics)". The main heading "THE GENE ONTOLOGY RESOURCE" is displayed in large white letters. Below this, a paragraph states: "The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life." A second paragraph follows: "The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research."

# Gene Ontology

Current release 2023-01-01: 43,248 GO terms | 7,503,460 annotations  
1,475,947 gene products | 5,267 species ([see statistics](#))

- GO is largest source of gene functional annotation
- Structured, controlled vocabulary of terms (and therefore gene sets)
- Manually annotated by a large consortium
- Data come from experimental and computational analyses

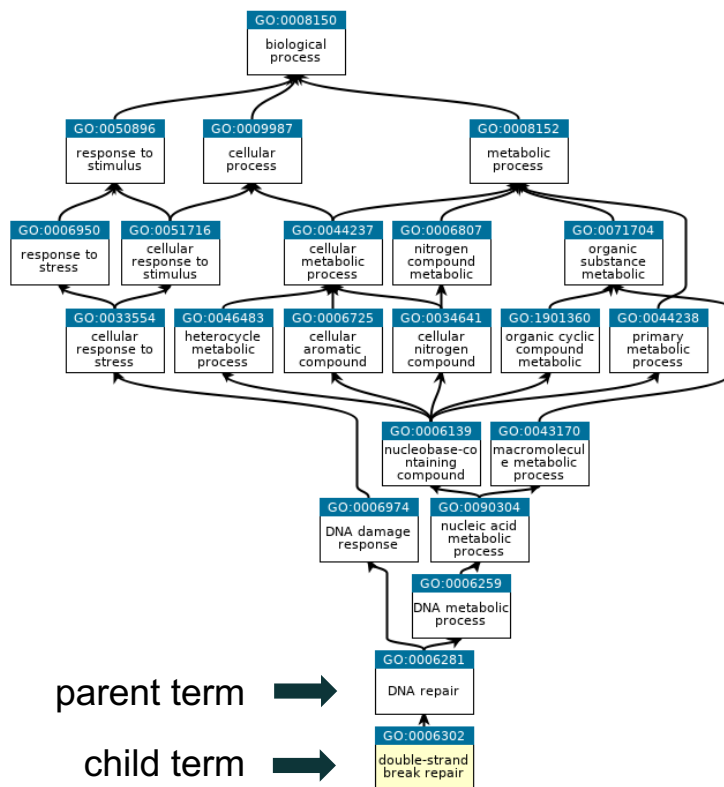
# GO ontologies

- Actually three separate ontologies:
  - **Molecular Function** – molecular level activities performed by gene products, e.g. *transporter activity* (broad) or *Toll-like receptor binding* (specific)
  - **Cellular Component** – the cellular location where a function is performed, e.g. *ribosome*
  - **Biological Process** – larger processes accomplished by multiple molecular activities, e.g. *DNA repair* (broad) or *pyrimidine nucleobase biosynthetic process* (specific)
- Generally, in functional enrichment analysis, “biological process” is most useful

# GO hierarchy

more genes

fewer genes



parent term

child term

root term

broad terms

specific terms

# BRCA2 example

## Gene: BRCA2 ENSG00000139618

Description	BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
Gene Synonyms	BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11
Location	Chromosome 13: 32,315,086-32,400,268 forward strand. GRCh38:CM000675.2
About this gene	This gene has 15 transcripts ( <a href="#">splice variants</a> ), <a href="#">173 orthologues</a> and is associated with <a href="#">120 phenotypes</a> .
Transcripts	<button>Show transcript table</button>

## GO: Molecular function ?

Show/hide columns (3 hidden)		Filter		
Accession	Term	Evidence	Annotation source	
<a href="#">GO:0002020</a>	protease binding	IPi	UniProt	
<a href="#">GO:0003677</a>	DNA binding	IEA	UniProt	
<a href="#">GO:0003697</a>	single-stranded DNA binding	IDA	UniProt	
<a href="#">GO:0005515</a>	protein binding	IPi	IntAct	
<a href="#">GO:0008022</a>	protein C-terminus binding	IDA	MGI	
<a href="#">GO:0010484</a>	H3 histone acetyltransferase activity	IDA	UniProt	
<a href="#">GO:0010485</a>	H4 histone acetyltransferase activity	IDA	UniProt	
<a href="#">GO:0042802</a>	identical protein binding	IPi	IntAct	
<a href="#">GO:0043015</a>	gamma-tubulin binding	IPi	UniProt	

# BRCA2 example

## Gene: BRCA2 ENSG00000139618

**Description** BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:[HGNC:1101](#)]

**Gene Synonyms** BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11

**Location** [Chromosome 13: 32,315,086-32,400,268](#) forward strand. GRCh38:CM000675.2

**About this gene** This gene has 15 transcripts ([splice variants](#)), [173 orthologues](#) and is associated with [120 phenotypes](#).

**Transcripts** [Show transcript table](#)

## GO: Molecular function ⓘ

Show/hide columns (3 hidden)				Filter	
Accession	Term	Evidence	Annotation source		
<a href="#">GO:0002020</a>	protease binding	IPI	UniProt		
<a href="#">GO:0003677</a>	DNA binding	IEA	UniProt		
<a href="#">GO:0003697</a>	single-stranded DNA binding	IDA	UniProt		
<a href="#">GO:0005515</a>	protein binding	IPI	IntAct		
<a href="#">GO:0008022</a>	protein C-terminus binding	IDA	MGI		
<a href="#">GO:0010484</a>	H3 histone acetyltransferase activity	IDA	UniProt		
<a href="#">GO:0010485</a>	H4 histone acetyltransferase activity	IDA	UniProt		
<a href="#">GO:0042802</a>	identical protein binding	IPI	IntAct		
<a href="#">GO:0043015</a>	gamma-tubulin binding	IPI	UniProt		

## Gene: BRCA2 ENSG00000139618

**Description** BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:[HGNC:1101](#)]

**Gene Synonyms** BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11

**Location** [Chromosome 13: 32,315,086-32,400,268](#) forward strand. GRCh38:CM000675.2

**About this gene** This gene has 15 transcripts ([splice variants](#)), [173 orthologues](#) and is associated with [120 phenotypes](#).

**Transcripts** [Show transcript table](#)

## GO: Cellular component ⓘ

Show All entries				Show/hide columns (3 hidden)	Filter	
Accession	Term	Evidence	Annotation source			
<a href="#">GO:0000152</a>	nuclear ubiquitin ligase complex	IDA	ComplexPortal			
<a href="#">GO:0000781</a>	chromosome, telomeric region	IDA	BHF-UCL			
<a href="#">GO:0000800</a>	lateral element	IDA	MGI			
<a href="#">GO:0005634</a>	nucleus	IDA, IEA	UniProt			
<a href="#">GO:0005654</a>	nucleoplasm	IDA	HPA			
<a href="#">GO:0005694</a>	chromosome	IEA	Ensembl			
<a href="#">GO:0005737</a>	cytoplasm	IEA	UniProt			
<a href="#">GO:0005813</a>	centrosome	IDA	UniProt			
<a href="#">GO:0005815</a>	microtubule organizing center	IEA	UniProt			
<a href="#">GO:0005829</a>	cytosol	IDA	HPA			
<a href="#">GO:0005856</a>	cytoskeleton	IEA	UniProt			
<a href="#">GO:0030141</a>	secretory granule	IDA	UniProt			
<a href="#">GO:0032991</a>	protein-containing complex	IDA	MGI			
<a href="#">GO:0033593</a>	BRCA2-MAGE-D1 complex	IDA	UniProt			
<a href="#">GO:1990391</a>	DNA repair complex	IPI	ComplexPortal			



# BRCA2 example

## Gene: BRCA2 ENSG00000139618

### Description

BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

### Gene Synonyms

BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11

### Location

Chromosome 13: 32,315,086-32,400,268 forward strand. GRCh38:CM000675.2

### About this gene

This gene has 15 transcripts (splice variants), 173 orthologues and is associated with 120 phenotypes.

### Transcripts

Show transcript table

## GO: Molecular function ?

Show/hide columns (3 hidden)

Accession	Term
<a href="#">GO:0002020</a>	protease binding
<a href="#">GO:0003677</a>	DNA binding
<a href="#">GO:0003697</a>	single-stranded DNA binding
<a href="#">GO:0005515</a>	protein binding
<a href="#">GO:0008022</a>	protein C-terminus binding
<a href="#">GO:0010484</a>	H3 histone acetyltransferase activity
<a href="#">GO:0010485</a>	H4 histone acetyltransferase activity
<a href="#">GO:0042802</a>	identical protein binding
<a href="#">GO:0043015</a>	gamma-tubulin binding

## Gene: BRCA2 ENSG00000139618

### Description

BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

### Gene Synonyms

BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11

### Location

Chromosome 13: 32,315,086-32,400,268 forward strand. GRCh38:CM000675.2

### About this gene

This gene has 15 transcripts (splice variants), 173 orthologues and is associated with 120 phenotypes.

### Transcripts

Show transcript table

## GO: Biological process ?

Show	All	entries	Show/hide columns (3 hidden)	Filter	
Accession	Term	Evidence	Annotation source		
<a href="#">GO:0000722</a>	telomere maintenance via recombination	IEA	Ensembl		
<a href="#">GO:0000724</a>	double-strand break repair via homologous recombination	IEA			
<a href="#">GO:0001556</a>	oocyte maturation	IEA	Ensembl		
<a href="#">GO:0001833</a>	inner cell mass cell proliferation	IEA	Ensembl		
<a href="#">GO:0006281</a>	DNA repair	IEA			
<a href="#">GO:0006289</a>	nucleotide-excision repair	IMP	UniProt		
<a href="#">GO:0006302</a>	double-strand break repair	IMP	UniProt		
<a href="#">GO:0006310</a>	DNA recombination	IEA	UniProt		
<a href="#">GO:0006355</a>	regulation of DNA-templated transcription	IBA	GO_Central		
<a href="#">GO:0006974</a>	cellular response to DNA damage stimulus	IEA	UniProt		
<a href="#">GO:0006978</a>	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	IEA	Ensembl		
<a href="#">GO:0007049</a>	cell cycle	IEA	UniProt		

ENSG00000139618

BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]

BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11

Chromosome 13: 32,315,086-32,400,268 forward strand.

GRCh38:CM000675.2

This gene has 15 transcripts (splice variants), 173 orthologues and is associated with 120 phenotypes.

Show transcript table

## Phenotype ?

Show/hide columns (3 hidden)

	Evidence	Annotation source
ar ubiquitin ligase complex	IDA	ComplexPortal
osome, telomeric region	IDA	BHF-UCL
element	IDA	MGI
us	IDA, IEA	UniProt
oplasm	IDA	HPA
osome	IEA	Ensembl
asm	IEA	UniProt
osome	IDA	UniProt
tubule organizing center	IEA	UniProt
pl	IDA	HPA
skeleton	IEA	UniProt
tory granule	IDA	UniProt
n-containing complex	IDA	MGI
2-MAGE-D1 complex	IDA	UniProt
repair complex	IPI	ComplexPortal

# Functional enrichment analysis

- How do we use all the existing annotation to interpret our gene list?
- Want to identify biological functions that are enriched in our gene list

# Testing for functional enrichment

20,000 genes assayed

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG000000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11
ENSDARG000000004748	3.380766458381055e-11
ENSDARG000000102808	3.652027442583773e-11
ENSDARG000000059294	3.68666857880831e-11
ENSDARG000000058731	7.374599119303466e-11
ENSDARG000000030896	1.0690865837043268e-10
ENSDARG000000117300	3.836221667423099e-10
ENSDARG000000102758	7.675777258160306e-10
ENSDARG000000004754	1.390777904676073e-9
ENSDARG000000099660	2.0021043563850804e-9
ENSDARG000000056615	2.032130501204873e-9
ENSDARG000000073820	3.749723988428957e-9
ENSDARG00000003570	4.027615069700461e-9
ENSDARG000000077799	7.492097636023496e-9
ENSDARG000000053836	7.754409869629003e-9
ENSDARG000000094678	7.81554132440011e-9
ENSDARG000000076914	2.8102379001185536e-8
ENSDARG000000037421	2.990985153534242e-8
ENSDARG000000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG000000018491	4.275633454932327e-8
ENSDARG000000109648	4.9384742480335793e-8
ENSDARG00000010231	5.822859503020341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.2504789486538527e-7
ENSDARG000000054196	1.356347303797386e-7
ENSDARG000000090548	1.8758159693858449e-7
ENSDARG000000104919	2.2367046789114162e-7
ENSDARG000000062783	3.710767241197336e-7
ENSDARG000000092227	4.383999193165627e-7
ENSDARG000000051888	4.6980261652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG0000000056196	4.7089485706046475e-7
ENSDARG000000018283	4.917667288531457e-7
ENSDARG000000105731	5.285812723974355e-7
ENSDARG000000004954	5.526970490118169e-7
ENSDARG000000025903	7.245067792685657e-7
ENSDARG000000025094	7.999474861543812e-7
ENSDARG000000100504	8.037125469772779e-7

Adjusted  
p-value  
< 0.05



500 significantly DE genes

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG000000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11

# Testing for functional enrichment

20,000 genes assayed

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11
ENSDARG000000004748	3.380766458381055e-11
ENSDARG000000102808	3.652027442583773e-11
ENSDARG000000059294	3.686666857880831e-11
ENSDARG000000058731	7.374599119303466e-11
ENSDARG000000030896	1.0690865837043268e-10
ENSDARG000000117300	3.836221667423099e-10
ENSDARG000000102758	7.67577258160306e-10
ENSDARG000000047454	1.390777904676073e-9
ENSDARG000000099660	2.0021043563850094e-9
ENSDARG000000056615	2.032130501204873e-9
ENSDARG000000073820	3.749723088428957e-9
ENSDARG00000003570	4.027615069700461e-9
ENSDARG000000077799	7.492097636023496e-9
ENSDARG000000053836	7.754409869629003e-9
ENSDARG000000094678	7.81554132440011e-9
ENSDARG000000076914	2.8102379001185536e-8
ENSDARG000000037421	2.990985153534242e-8
ENSDARG000000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG00000018491	4.275633454932327e-8
ENSDARG000000109648	4.9384742480335793e-8
ENSDARG00000010231	5.822859503020341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.2504789486538527e-7
ENSDARG000000054196	1.356347303797386e-7
ENSDARG000000090548	1.8758159693858449e-7
ENSDARG000000104919	2.2367046789114162e-7
ENSDARG000000052787	3.7107672341197336e-7
ENSDARG000000092227	4.382399103165207e-7
ENSDARG000000051888	4.6980261652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG000000056196	4.7089485706046475e-7
ENSDARG00000018283	4.917667288531457e-7
ENSDARG000000105731	5.285812723974355e-7
ENSDARG000000004954	5.526970490118169e-7
ENSDARG000000025903	7.245067792685657e-7
ENSDARG000000025094	7.999474861543812e-7
ENSDARG000000100504	8.037125469772779e-7

Adjusted  
p-value  
< 0.05



500 significantly DE genes

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11



200 genes annotated  
to DNA repair

200/500 = **40%**

(300 not annotated to  
DNA repair)

2000 genes annotated to  
function (e.g. DNA repair)



2000/20000 = **10%**

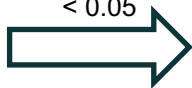
(18,000 not annotated to  
DNA repair)

# Testing for functional enrichment

20,000 genes assayed

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11
ENSDARG000000004748	3.380766458381055e-11
ENSDARG000000102808	3.652027442583773e-11
ENSDARG000000059294	3.686666857880831e-11
ENSDARG000000058731	7.374599119303466e-11
ENSDARG000000030896	1.0690865837043268e-10
ENSDARG000000117300	3.836221667423099e-10
ENSDARG000000102758	7.67577258160306e-10
ENSDARG00000004754	1.380777904676073e-9
ENSDARG000000099660	2.0021043563850804e-9
ENSDARG000000056615	2.032130501204873e-9
ENSDARG000000073820	3.749723088428957e-9
ENSDARG00000003570	4.027615069700461e-9
ENSDARG000000077799	7.492097636023496e-9
ENSDARG000000053836	7.754409869629003e-9
ENSDARG000000094678	7.81554132440011e-9
ENSDARG000000076914	2.8102379001185536e-8
ENSDARG000000037421	2.990985153534242e-8
ENSDARG00000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG00000018491	4.275633454932327e-8
ENSDARG000000109648	4.938474248035793e-8
ENSDARG00000010231	5.822859503020341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.2504789486538527e-7
ENSDARG000000054196	1.356347303797386e-7
ENSDARG000000090548	1.8758159693858449e-7
ENSDARG000000104919	2.2367046789114162e-7
ENSDARG000000062783	3.710767241197336e-7
ENSDARG000000092227	4.3837909103165207e-7
ENSDARG000000051888	4.6980261652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG000000056196	4.7089485706046475e-7
ENSDARG00000018283	4.917667288531457e-7
ENSDARG000000105731	5.285812723974355e-7
ENSDARG000000004954	5.526970490118169e-7
ENSDARG000000025903	7.245067792685657e-7
ENSDARG000000025094	7.999474861543812e-7
ENSDARG000000100504	8.037125469772779e-7

Adjusted  
p-value  
< 0.05



500 significantly DE genes

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626339e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG000000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11

200 genes annotated  
to DNA repair

200/500 = **40%**

(300 not annotated to  
DNA repair)

2000 genes annotated to  
function (e.g. DNA repair)

2000/20000 = **10%**

(18,000 not annotated to  
DNA repair)

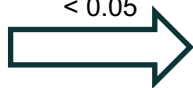
Is seeing 200 DNA repair  
genes significantly differentially  
expressed more than we would  
expect by chance?

# Testing for functional enrichment

20,000 genes assayed

```
Gene adjp
ENSDARG000000041294 1.0867269119101765e-32
ENSDARG000000060498 1.2517176061993635e-21
ENSDARG000000031683 2.364463411626339e-21
ENSDARG000000077982 2.9548243274497384e-14
ENSDARG000000070480 5.735034663692255e-14
ENSDARG000000007769 1.5678213864306653e-13
ENSDARG000000102435 1.9075360227976884e-13
ENSDARG000000101482 2.6986616800262944e-13
ENSDARG000000034503 5.567338300152267e-13
ENSDARG00000013670 2.3318547986985375e-11
ENSDARG000000039490 3.047413661053603e-11
ENSDARG000000004748 3.380766458381055e-11
ENSDARG000000102808 3.652027442583773e-11
ENSDARG000000059294 3.68666857880831e-11
ENSDARG000000058731 7.374599119303466e-11
ENSDARG000000030896 1.0690865837043268e-10
ENSDARG000000117300 3.836221667423099e-10
ENSDARG000000102758 7.675777258160306e-10
ENSDARG000000004754 1.390777904676073e-9
ENSDARG000000099609 2.0021043563850804e-9
ENSDARG000000056615 2.032130501204873e-9
ENSDARG000000073820 3.749723088428957e-9
ENSDARG00000003570 4.027615069700461e-9
ENSDARG000000077799 7.492097636023496e-9
ENSDARG000000053836 7.754409869629003e-9
ENSDARG000000094678 7.81554132440011e-9
ENSDARG000000076914 2.8102379001185536e-8
ENSDARG000000037421 2.990985153534242e-8
ENSDARG00000014340 3.505316564907291e-8
ENSDARG000000104773 4.1069391754303735e-8
ENSDARG000000105749 4.1069391754303735e-8
ENSDARG00000018491 4.275633454932327e-8
ENSDARG000000109648 4.938474248035793e-8
ENSDARG00000010231 5.82285993020341e-8
ENSDARG000000039142 7.665381608460826e-8
ENSDARG000000104672 1.2504789486538527e-7
ENSDARG000000054196 1.356347303797386e-7
ENSDARG000000090548 1.8758159693858449e-7
ENSDARG000000104919 2.2367046789114162e-7
ENSDARG000000062703 3.7107672314107336e-7
ENSDARG000000092227 4.3823999193165207e-7
ENSDARG000000051888 4.6980261652096274e-7
ENSDARG000000077169 4.7089485706046475e-7
ENSDARG000000056196 4.7089485706046475e-7
ENSDARG00000018283 4.917667288531457e-7
ENSDARG000000105731 5.285812723974355e-7
ENSDARG000000004954 5.526970490118169e-7
ENSDARG000000025903 7.245067792685657e-7
ENSDARG000000025094 7.999474861543812e-7
ENSDARG000000100504 8.037125469772779e-7
```

Adjusted  
p-value  
< 0.05



500 significantly DE genes

```
Gene adjp
ENSDARG000000041294 1.0867269119101765e-32
ENSDARG000000060498 1.2517176061993635e-21
ENSDARG000000031683 2.364463411626339e-21
ENSDARG000000077982 2.9548243274497384e-14
ENSDARG000000070480 5.735034663692255e-14
ENSDARG000000007769 1.5678213864306653e-13
ENSDARG000000102435 1.9075360227976884e-13
ENSDARG000000101482 2.6986616800262944e-13
ENSDARG000000034503 5.567338300152267e-13
ENSDARG00000013670 2.3318547986985375e-11
ENSDARG000000039490 3.047413661053603e-11
```

200 genes annotated  
to DNA repair

200/500 = **40%**

(300 not annotated to  
DNA repair)

2000 genes annotated  
to function (e.g. DNA

2000/20000 = **10%**

(18,000 not annotated  
to DNA repair)

	DE	Not DE	Total
Annotated to DNA repair	200	1800	2000
Not annotated to DNA repair	300	17700	18000
<b>Total</b>	<b>500</b>	<b>19500</b>	<b>20000</b>

# Hypergeometric test

	DE	Not DE	Total
Annotated to DNA repair	200	1800	2000
Not annotated to DNA repair	300	17700	18000
<b>Total</b>	<b>500</b>	<b>19500</b>	<b>20000</b>

```
> m <- 20000 # Total genes
> n <- 500    # Number of DE genes
> mt <- 2000  # Number of annotated genes
> nt <- 200   # Number of annotated DE genes
> phyper(nt - 1, mt, m - mt, n, lower.tail=FALSE)
[1] 1.65531e-72
```

Use the hypergeometric test to calculate the probability of having 200 or more DE annotated genes when 2000 of the 20,000 total genes are annotated

$$P(\sigma_t \geq n_t) = \sum_{k=n_t}^{\min(m_t, n)} \frac{\binom{m_t}{k} \binom{m-m_t}{n-k}}{\binom{m}{n}}$$

# Multiple testing correction

- In reality, won't just be doing one test
- Want to test all (or a lot) of the GO terms and other functional gene sets
- Leads to problem of **multiple testing**
- If you test 10,000 GO terms with a significance threshold of  $< 0.05$  then you expect 500 terms to be significant simply by chance
- Need to correct for multiple testing:
  - Bonferroni
  - Benjamini–Hochberg



# Bonferroni correction

- Bonferroni is easiest to understand and most conservative
- Simply multiply all p-values by the number of tests (i.e. functional gene sets)
- Get adjusted p-values

Gene	pval	adjp
ENSDARG00000041294	4.904e-37	4.904e-36
ENSDARG00000027744	0.0299254	0.299254
ENSDARG00000100466	0.0412905	0.412905
ENSDARG00000036912	0.0638413	0.638413
ENSDARG00000040553	0.12413	1.00000
ENSDARG00000002016	0.141297	1.00000
ENSDARG00000035256	0.150329	1.00000
ENSDARG00000092381	0.221648	1.00000
ENSDARG00000101125	0.224041	1.00000
ENSDARG00000041223	0.224123	1.00000

# Benjamini–Hochberg correction

- Benjamini–Hochberg is less conservative and assumes that all tests are statistically independent
- Not true – many functional gene sets overlap:
  - e.g. GO terms are hierarchical so a term's annotations are a subset of their parental annotations
  - e.g. similar pathways can appear in KEGG and WikiPathways
  - e.g. some genes are co-expressed
- Nevertheless, BH is widely and successfully used
- Although Wijesooriya *et al.* (2022) found that 43% of papers surveyed failed to do multiple testing correction:  
[doi.org/10.1371/journal.pcbi.1009935](https://doi.org/10.1371/journal.pcbi.1009935)

# Background gene set

- Important to choose appropriate background gene set
- Wijesooriya *et al.* (2022) found that only 4% of papers used an appropriate background (although most failed to specify what background was used): [doi.org/10.1371/journal.pcbi.1009935](https://doi.org/10.1371/journal.pcbi.1009935)
- Best to choose all genes that could have been captured in your experiment
- Examples:
  - All genes
  - All genes with non-zero total read count in DESeq2
  - All genes that pass DESeq2 independent filtering
  - All genes expressed in a particular tissue
  - All genes with annotations

# Other methods

- Functional enrichment analysis (or over-representation analysis) is just one method
- Other methods and tests are available, e.g.
  - GSEA (gene set enrichment analysis)
  - Binomial test
- Concentrating on functional enrichment analysis because most widely used and most tools available

# Advantages of functional enrichment analysis

- Improves statistical power as you effectively sum up counts from the multiple genes in a functional gene set
- Improves statistical power as there are usually fewer functional annotations than genes, so less multiple testing correction is needed
- Results are easier to interpret because they are familiar concepts like “DNA repair” rather than obscure gene names
- Diverse data (e.g. RNA-seq, proteomics) can be integrated because they map to common terms/pathways
- Results may be more comparable to related data because results are projected to a smaller set of functional annotations

# Disadvantages of functional enrichment analysis

- Terms or pathways with few genes are unlikely to ever be enriched
- Hypergeometric test is more likely to identify larger functional gene sets (e.g. pathways with many genes) as significant
- Genes with multiple functions can lead to enrichment of multiple terms/pathways, some of which aren't relevant
- Databases are (obviously) biased towards genes with annotation so unannotated genes (e.g. many non-coding RNA genes) are invisible to functional enrichment analysis

# Recommendations based on disadvantages

- For human RNA-seq data, consider excluding functional gene sets with  $< 10$  genes and  $> 500$  genes
- Former are unlikely to ever be significant and latter are too likely to be significant and will often be better represented by other more specific terms/pathways
- Always think about your own experiment:
  - e.g. is apoptosis enrichment expected or a symptom of a problem during sample preparation

# Quiz!

- Quiz on Mentimeter ([www.menti.com](https://www.menti.com))



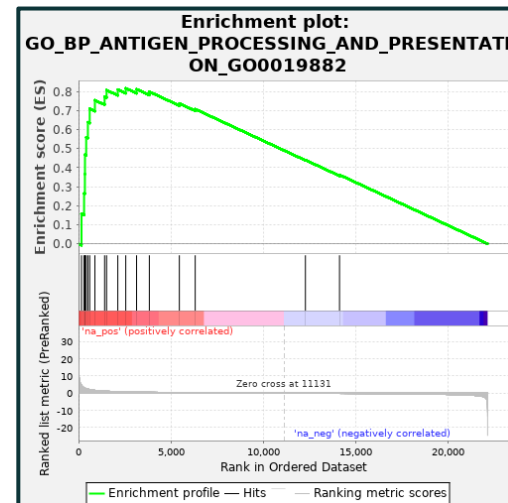
# Functional enrichment tools

- Many, many functional enrichment analysis tools exist
- Many are created, published and then never updated
- Best to choose a well used tool
- Using g:Profiler because:
  - Consistently and regularly updated over many years
  - Easy to use
  - Free
  - Well documented
  - Has advanced features, like simultaneous analysis of multiple lists
  - Has web interface but also an API with supported R and Python packages
  - Covers over 800 species/strains



# Other functional enrichment tools

- Other tools are available (and good):
  - Enrichr ([maayanlab.cloud/Enrichr/](https://maayanlab.cloud/Enrichr/)):
    - Web-based
    - Similar to g:Profiler
    - Only human, mouse, fly, yeast, worm and zebrafish
  - GSEA ([www.gsea-msigdb.org/gsea/](https://www.gsea-msigdb.org/gsea/)):
    - Desktop software
    - Implements GSEA method
    - Works on whole genome ranked gene lists
    - Looks for gene sets enriched at top or bottom of your ranked list
    - p-values computed by permutating ranked lists



# g:Profiler

- g:Profiler uses Ensembl as its primary data source (specifically, BioMart)
- Tracks Ensembl release schedule (every three or four months) but with delay of weeks or months
- g:Profiler beta covering Ensembl 108 (released in October) came out yesterday, but Ensembl 109 came out two weeks ago
- Recommend using Ensembl IDs as input, but not essential

# g:Profiler

g:Profiler

[News](#)

[Archives](#)

[Beta](#)

[API](#)

[R client](#)

[FAQ](#)

[Docs](#)

[Contact](#)

[Cite g:Profiler](#)

[Services using g:P](#)

[List of organisms](#)



**g:GOS**

Functional profiling

**g:Convert**

Gene ID conversion

**g:Orth**

Orthology search

**g:SNPense**

SNP id to gene name

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

## Options

Organism: ?

Homo sapiens (Human)

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼

Run query

random example

mixed query example

# g:Profiler – four tools

g:Profiler

[News](#)[Archives](#)[Beta](#)[API](#)[R client](#)[FAQ](#)[Docs](#)[Contact](#)[Cite g:Profiler](#)[Services using g:P](#)[List of organisms](#)

**g:GOS**  
Functional profiling

**g:Convert**  
Gene ID conversion

**g:Orth**  
Orthology search

**g:SNPense**  
SNP id to gene name

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes

Run query

random example

mixed query example

Options

Organism: 

Homo sapiens (Human)

☐ Ordered query

☐ Run as multiquery

Advanced options

Data sources

Bring your data (Custom GMT)

# g:Profiler – gene list

g:Profiler

NewsArchivesBetaAPIR clientFAQDocsContactCite g:ProfilerServices using g:PList of organisms

g:GOSt

Functional profiling

g:Convert

Gene ID conversion

g:Orth

Orthology search

g:SNPense

SNP id to gene name

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

Run query

random example

mixed query example

Options

Organism

Homo s

105 identifiers recognised for human

82 for mouse; 92 for zebrafish

☐ Order

☐ Run as

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼

# g:Profiler – options

g:Profiler

NewsArchivesBetaAPIR clientFAQDocsContactCite g:ProfilerServices using g:PList of organisms

g:GOSt

Functional profiling

g:Convert

Gene ID conversion

g:Orth

Orthology search

g:SNPense

SNP id to gene name

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

821 species/strains in current release

(849 in next release/beta)

Run query

random example

mixed query example

Options

Organism: ?

Homo sapiens (Human)

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ▾

Data sources ▾

Bring your data (Custom GMT) ▾

# g:Profiler – advanced options

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

g:SCS – “Set Counts and Sizes”

Accounts for hierarchical nature of GO

Less conservative than Bonferroni but more conservative than Benjamini-Hochberg

## Options

Organism: ?

Homo sapiens (Human) ▼

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ^

☐ All results ?

☐ Measure underrepresentation ?

☐ No evidence codes ?

Statistical domain scope ?

Only annotated genes ▼

Significance threshold ?

g:SCS threshold ▼

User threshold ?

0.05

Numeric IDs treated as ?

ENTREZGENE\_ACC ▼



# g:Profiler – data sources

9 data sources

(or 11 if count GO as  
three separate sources)

All 9 not available for all  
species



**Data sources** ^

[Show data versions](#)

**Gene Ontology**

☒ GO molecular function

☒ GO cellular component

☒ GO biological process

☐ No electronic GO annotations ?

**biological pathways**

☒ KEGG

☒ Reactome

☒ WikiPathways

**regulatory motifs in DNA**

☒ TRANSFAC

☒ miRTarBase

**protein databases**

☒ Human Protein Atlas

☒ CORUM

**Human phenotype ontology**

☒ HP

[name.gmt zip](#) [combined name.gmt](#)

[ENSG.gmt zip](#) [combined ENSG.gmt](#)



Can exclude GO IEA  
evidence term (inferred  
from electronic  
annotation)

But often as reliable as  
human annotation  
(Škunca *et al.* 2012)

Suggest running with  
and without if using  
human or model  
organisms

# g:Profiler – bring your data

Query

Upload query

Upload bed file

Input is whitespace-separated list of genes ?

Run query

random example

mixed query example

Options

Organism: ?  
Homo sapiens (Human) ▼

☐ Ordered query ?  
☐ Run as multiquery ?

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ^



Drag GMT file (or ZIP archive with GMT files) here to begin  
or click to browse

File name used\*  
not selected ▼

or insert token:  
 +

\* If changed query will be run across the selected GMT.  
For public GMT sources see link

# g:Profiler – documentation

 [News](#) [Archives](#) [Beta](#) [API](#) [R client](#) [FAQ](#) [Docs](#) [Contact](#) [Cite g:Profiler](#) [Services using g:P](#) [List of organisms](#) 

**g:GOS**  
Functional profiling

**g:Convert**  
Gene ID conversion

**g:Orth**  
Orthology search

**g:SNPense**  
SNP id to gene name

## Welcome to g:Profiler

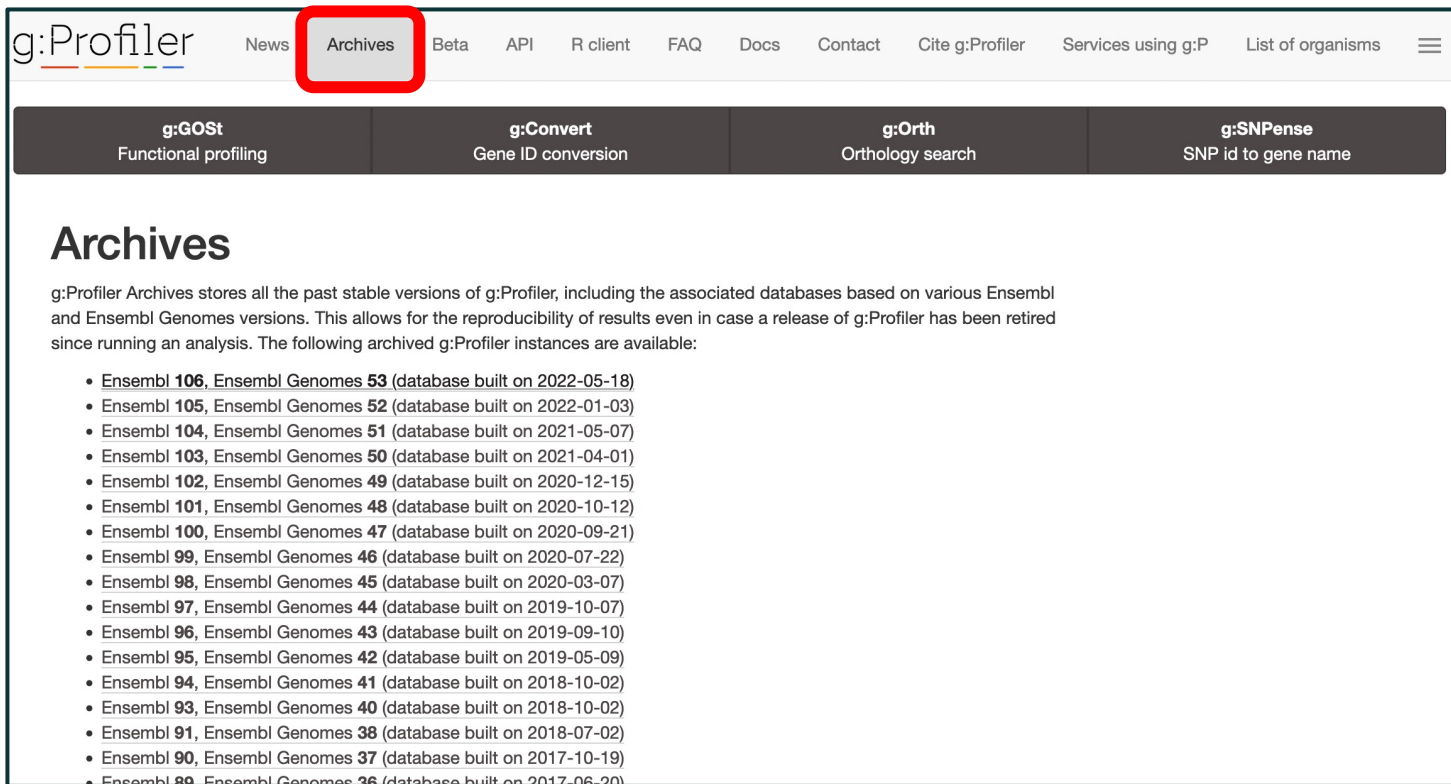
g:Profiler is a public web server for characterising and manipulating gene lists. g:Profiler has a simple user-friendly web interface with powerful visualisations and is currently available for 400+ species, including mammals, plants, fungi, insects from Ensembl and Ensembl Genomes. g:Profiler is updated approximately in every three months and follows quarterly releases of Ensembl databases. g:Profiler tool set consists of the following tools:

- **g:GOS**, the core of the g:Profiler, performs statistical enrichment analysis to provide interpretation to user-provided gene lists. The gene lists can be either flat or ordered gene lists. We accept majority of the identifier types, chromosomal regions and term IDs as input. We provide data from multiple sources of functional evidence, including Gene Ontology terms, biological pathways, regulatory motifs of transcription factors and microRNAs, human disease annotations and protein-protein interactions.
- **g:Convert** is a gene identifier conversion tool. It uses information in Ensembl databases to handle hundreds of types of IDs for genes, proteins, transcripts, microarray probesets, etc, for many species, experimental platforms and biological databases. g:Convert is flexible: it accepts a mixed list of IDs and recognises their types automatically. It can also serve as a service to get all genes belonging to a particular functional category.
- **g:Orth** is a tool for mapping homologous genes across related organisms based on Ensembl data. Given a selected target organism, g:Orth retrieves the genes of the target organism that are similar in sequence to the initial genes in the input.
- **g:SNPense** is a tool for mapping human single nucleotide polymorphisms (SNP) to gene names, chromosomal locations and variant consequence terms from Sequence Ontology.

### Contents

- Welcome to g:Profiler
  - About g:Profiler
  - Publications and theses
  - Funding
  - Tech notes
  - Support
- g:GOS
  - Using g:GOS
  - Examples
- g:Convert
  - Using g:Convert
  - Examples
- g:Orth
  - Using g:Orth
  - Examples
- g:SNPense

# g:Profiler – archives



**g:Profiler** News **Archives** Beta API R client FAQ Docs Contact Cite g:Profiler Services using g:P List of organisms

**g:GOST**  
Functional profiling

**g:Convert**  
Gene ID conversion

**g:Orth**  
Orthology search

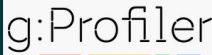

**g:SNPense**  
SNP id to gene name

## Archives

g:Profiler Archives stores all the past stable versions of g:Profiler, including the associated databases based on various Ensembl and Ensembl Genomes versions. This allows for the reproducibility of results even in case a release of g:Profiler has been retired since running an analysis. The following archived g:Profiler instances are available:

- [Ensembl 106](#), [Ensembl Genomes 53](#) (database built on 2022-05-18)
- [Ensembl 105](#), [Ensembl Genomes 52](#) (database built on 2022-01-03)
- [Ensembl 104](#), [Ensembl Genomes 51](#) (database built on 2021-05-07)
- [Ensembl 103](#), [Ensembl Genomes 50](#) (database built on 2021-04-01)
- [Ensembl 102](#), [Ensembl Genomes 49](#) (database built on 2020-12-15)
- [Ensembl 101](#), [Ensembl Genomes 48](#) (database built on 2020-10-12)
- [Ensembl 100](#), [Ensembl Genomes 47](#) (database built on 2020-09-21)
- [Ensembl 99](#), [Ensembl Genomes 46](#) (database built on 2020-07-22)
- [Ensembl 98](#), [Ensembl Genomes 45](#) (database built on 2020-03-07)
- [Ensembl 97](#), [Ensembl Genomes 44](#) (database built on 2019-10-07)
- [Ensembl 96](#), [Ensembl Genomes 43](#) (database built on 2019-09-10)
- [Ensembl 95](#), [Ensembl Genomes 42](#) (database built on 2019-05-09)
- [Ensembl 94](#), [Ensembl Genomes 41](#) (database built on 2018-10-02)
- [Ensembl 93](#), [Ensembl Genomes 40](#) (database built on 2018-10-02)
- [Ensembl 91](#), [Ensembl Genomes 38](#) (database built on 2018-07-02)
- [Ensembl 90](#), [Ensembl Genomes 37](#) (database built on 2017-10-19)
- [Ensembl 89](#), [Ensembl Genomes 36](#) (database built on 2017-06-20)

# g:Profiler – API and libraries

 [News](#) [Archives](#) [Beta](#) [API](#) [R client](#) [FAQ](#) [Docs](#) [Contact](#) [Cite g:Profiler](#) [Services using g:P](#) [List of organisms](#) 

**g:GOST**  
Functional profiling

**g:Convert**  
Gene ID conversion

**g:Orth**  
Orthology search

**g:SNPense**  
SNP id to gene name

## g:Profiler client libraries

### R client

g:Profiler has an up-to-date R client library `gprofiler2` available from CRAN or conda-forge. For more documentation see [the help page](#)

### Python client

g:Profiler has an up-to-date python client library `gprofiler-official` available from PyPI or conda-forge. For more documentation see [the package description](#)

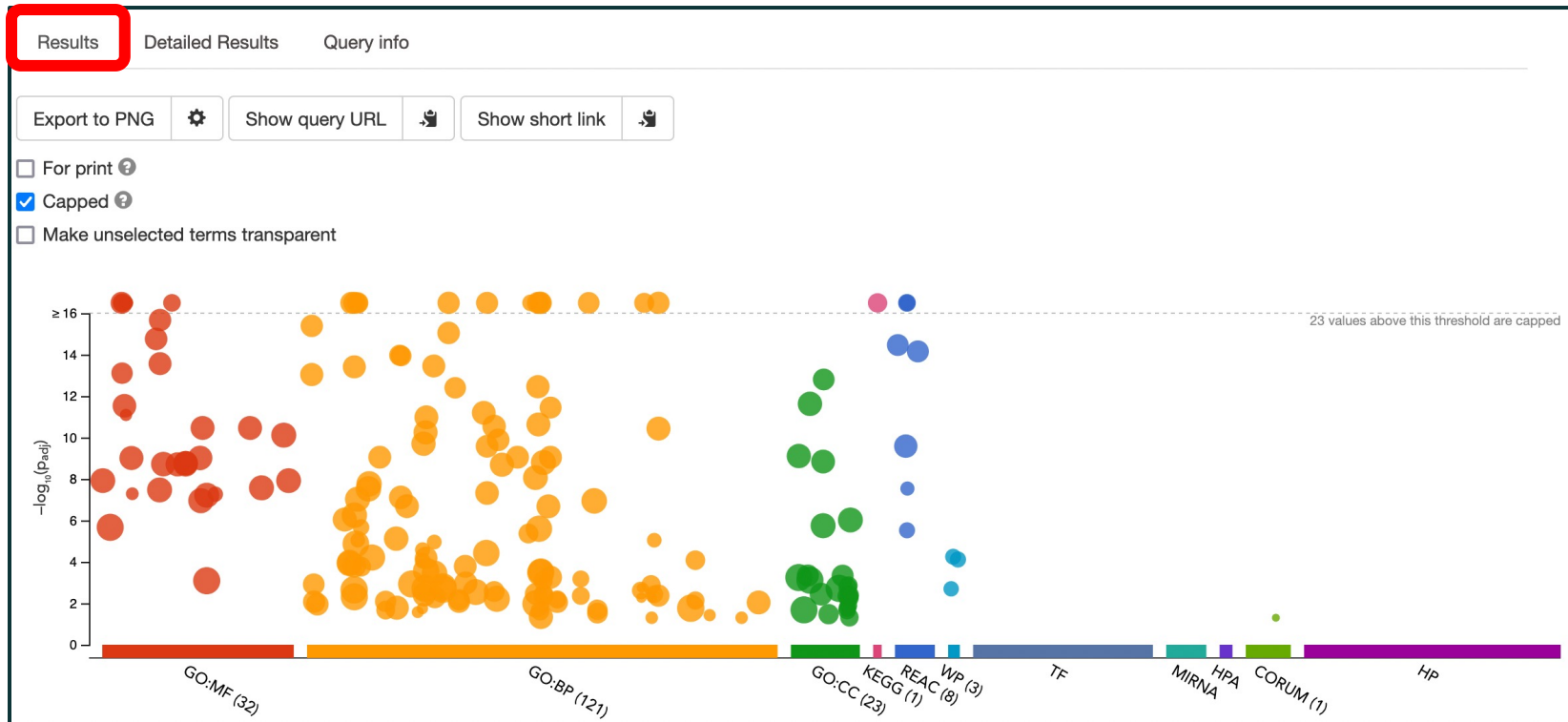
### g:Profiler API

g:Profiler requests are generally made as POST requests with a JSON body and they return JSON output.

#### Contents

- R client
- Python client
- g:GOST
  - g:GOST query result fields
  - Simple python example
  - Simple CURL example
  - Python example with more parameters set to non-default values
- g:Convert
  - Simple python example
  - Simple CURL example
- g:Orth
  - Simple python example
  - Simple CURL example
- g:SNPense
  - Simple python example
  - Simple CURL example

# g:Profiler – results



## g:Profiler – detailed results

[illegible]

# g:Profiler – beta

g:Profiler<sup>β</sup>

NewsArchivesStableAPIR clientFAQDocsContactCite g:ProfilerServices using g:PList of organisms

g:GOST  
Functional profiling

g:Convert  
Gene ID conversion

g:Orth  
Orthology search

g:SNPense  
SNP id to gene name

QueryUpload queryUpload bed file

Input is whitespace-separated list of genes ?

Options

Organism: ?  
Homo sapiens (Human)

☒ Highlight driving GO terms ?

☐ Ordered query ?

☐ Run as multiquery ?

Advanced options ▼

Data sources ▼

Bring your data (Custom GMT) ▼

Run query

random example

mixed query example



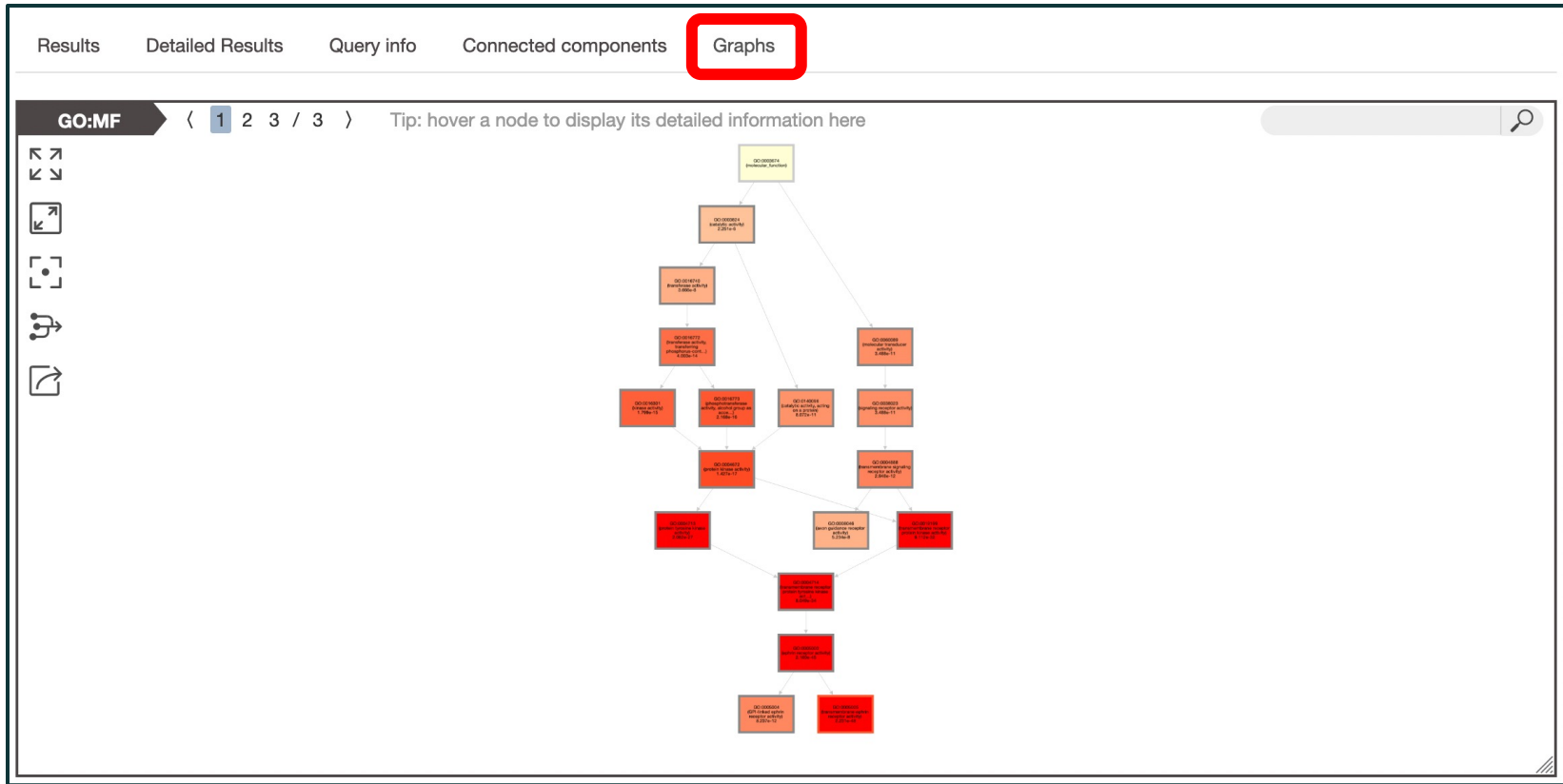
# g:Profiler – beta results



# g:Profiler – beta connected components

Results	Detailed Results	Query info	Connected components	Graphs
source	term id	term name	T	Padj (query_1)
GO:MF	GO:0005005	transmembrane-ephrin receptor activity	16	$2.231 \times 10^{-48}$
GO:MF	GO:0005003	ephrin receptor activity	19	$2.160 \times 10^{-45}$
GO:MF	GO:0004714	transmembrane receptor protein tyrosine kinas...	63	$8.049 \times 10^{-34}$
GO:MF	GO:0019199	transmembrane receptor protein kinase activity	82	$9.112 \times 10^{-32}$
GO:MF	GO:0004713	protein tyrosine kinase activity	147	$2.082 \times 10^{-27}$
GO:MF	GO:0004672	protein kinase activity	587	$1.427 \times 10^{-17}$
GO:MF	GO:0016773	phosphotransferase activity, alcohol group as ...	696	$2.168 \times 10^{-16}$
GO:MF	GO:0016301	kinase activity	795	$1.799 \times 10^{-15}$
GO:MF	GO:0016772	transferase activity, transferring phosphorus-c...	967	$4.003 \times 10^{-14}$
GO:MF	GO:0004888	transmembrane signaling receptor activity	1271	$2.948 \times 10^{-12}$
GO:MF	GO:0005004	GPI-linked ephrin receptor activity	7	$8.237 \times 10^{-12}$
GO:MF	GO:0060089	molecular transducer activity	1489	$3.488 \times 10^{-11}$
GO:MF	GO:0038023	signaling receptor activity	1489	$3.488 \times 10^{-11}$
GO:MF	GO:0140096	catalytic activity, acting on a protein	2335	$8.072 \times 10^{-11}$
GO:MF	GO:0016740	transferase activity	2340	$3.666 \times 10^{-8}$
GO:MF	GO:0008046	axon guidance receptor activity	9	$5.234 \times 10^{-8}$
GO:MF	GO:0003824	catalytic activity	5691	$2.251 \times 10^{-6}$
source	term id	term name	T	Padj (query_1)
GO:MF	GO:0035639	purine ribonucleoside triphosphate binding	1848	$9.889 \times 10^{-10}$
GO:MF	GO:0005524	ATP binding	1500	$9.902 \times 10^{-10}$
GO:MF	GO:0032555	purine ribonucleotide binding	1917	$1.739 \times 10^{-9}$
GO:MF	GO:0032559	adenyl ribonucleotide binding	1563	$1.795 \times 10^{-9}$

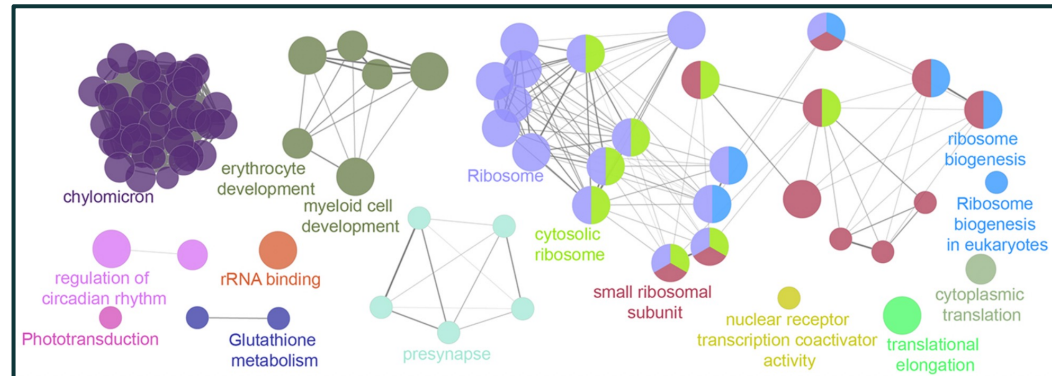
# gProfiler – beta graphs



# Summarising functional enrichment

- Functional enrichment analysis (hopefully) summarises a gene list into something shorter and more comprehensible
- But what if the list of functional enrichments is also long and/or repetitive?
- The connected components functionality is an attempt to solve that problem
- Other methods:

- Cytoscape / EnrichmentMap
- Cytoscape / ClueGO



## g:Profiler live demo!

- [biit.cs.ut.ee/gprofiler/](http://biit.cs.ut.ee/gprofiler/)

# Exercises (plus data and slides)

- Exercises are available from:

[rnaseq2023.buschlab.org](https://rnaseq2023.buschlab.org)

- Plus data for exercises and these slides
- Everything also available on penelopeprime