

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

14th Feb 2024



Me

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

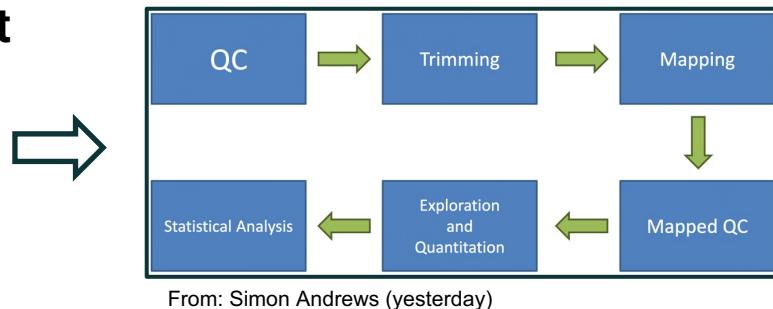
The screenshot shows the homepage of the Busch Lab website. The header features a circular logo with the text "Busch Lab" and three stylized nodes connected by lines. Below the header is a navigation menu with links: Home, Contact, News, Publications, Research, Resources, Software, Team, and ZMP. The main content area contains a large word cloud centered around "gene regulatory networks". Other prominent words include "zebrafish", "genetic screen", "RNA-seq", "mutants", "development", "chromatin", and "regeneration". A descriptive text below the word cloud states: "The Busch Lab is based in London, UK and our research focusses on investigating *in vivo* gene regulatory responses to challenges such as genetic mutations, chromatin disruption and infection." At the bottom right is a "Follow" button.

Questions

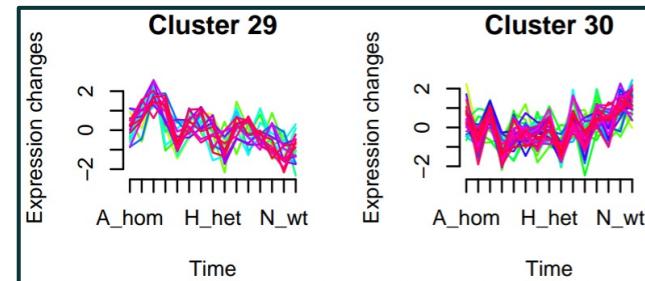
- For urgent questions, either:
 - Use Zoom's Chat
 - Unmute and ask
- If you can wait, then ask on Slack and I'll answer later

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?

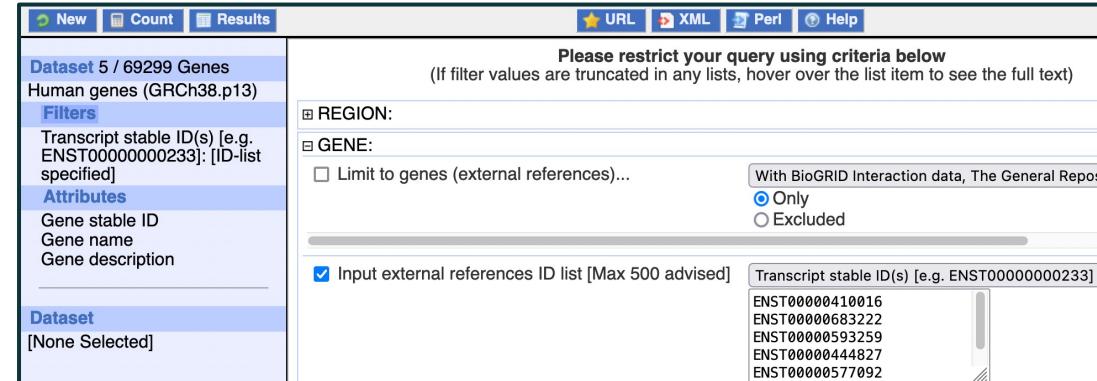
- Gene list can be:
 - Unranked (e.g. genes with somatic mutations in cancer sample)
 - Ranked (e.g. sensitivity in a CRISPR screen)
- RNA-seq differential expression analysis produces ranked lists
- Ranked lists are ordered by a score or metric:
 - e.g. adjusted p-value
 - e.g. \log_2 fold change
-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



The screenshot shows the BioMart interface for searching genes. At the top, there are tabs for 'New', 'Count', and 'Results'. To the right are links for 'URL', 'XML', 'Perl', and 'Help'. Below these are sections for '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), and 'Dataset' ([None Selected]). On the right, a message says 'Please restrict your query using criteria below' and '(If filter values are truncated in any lists, hover over the list item to see the full text)'. It includes sections for 'REGION:' and 'GENE:', with options like 'Limit to genes (external references)...' (radio buttons for 'Only' and 'Excluded'), 'With BioGRID Interaction data, The General Repository...', and 'Input external references ID list [Max 500 advised]' (checkbox checked). A list of transcript stable IDs is shown on the right: 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
ENSDARG0000007769	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
ENSDARG0000034503	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	Chr	Start	End	Name	Description
ENSDARG00000041294	4.904002310063973e-37	1.0867269119101765e-32	1.5709251030700861	3	62161184	62169060	noxo1a	NADPH oxidase organizer 1a

ENSDARG00000041294	4.904002310063973e-37	1.0867269119101765e-32	1.5709251030700861	3	62161184	62169060	noxo1a	NADPH oxidase organizer 1a
ENSDARG0000006498	1.1297093030865851e-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
ENSDARG0000007769	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
ENSDARG0000034503	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					
ENSDARG0000034503	2.261103100242347e-16	5.567338300152267e-13	0.6082489350504545					

Look up genes in databases

GENE

noxo1a

ID ZDB-GENE-030131-9700

Name *NADPH oxidase organizer 1a*

Symbol noxo1a Nomenclature History

Previous Names noxo1, cb18 (1), sb:cb18, SNX28b (1), wu:fd09d09, zgc:152911 (1)

Type protein_coding_gene ↗

Location Chr: 3 [Mapping Details/Browsers](#)

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 NOXO1 (NADPH oxidase organizer 1).

Genome Resources [Alliance](#) (1), [Gene:572245](#) (1), [Ensembl\(GRCz11\):ENSDARG00000041294](#) (3)

Note None

Comparative Information 

Look up genes in databases

GENE

 noxo GENE

NOXO1

ID	
Name	
Symbol	
Previous N	Species <i>Homo sapiens</i>
Type	Symbol NOXO1
Location	Name NADPH oxidase organizer 1
Description	Synonyms MGC20258 NADPH oxidase regulatory protein ▼ Show All 12
Genome Re	
Note	Biotype protein coding gene
Comparati	Automated Description ? Enables enzyme binding activity. Involved in extracellular matrix disassembly. Part of NADPH oxidase complex.
Information	
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	ENSEMBL:ENSG00000196408 NCBI_Gene:124056 ▼ Show All 4
Additional Information	Literature

Look up genes in databases

GENE

noxo GENE

ID
Name
Symbol
Previous N
Type
Location
Description
Genome Re
Note
Comparativ
Information

Species
Symbol
Name
Synonyms
Biotype
Automated

RGD Descrip
Cross Referenc

Additional Information Literature ↗

NOXO1

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

Look up genes in databases

GENE

noxo GENE

NOXO1

ID	Name	Symbol	Species	Previous Name	Type	Location	Description	Synonyms	Genome Reference	Note	Comparative Information	RGD Description	Cross References	Additional Information

Summaries for NOXO1 Gene

Entrez Gene Summary [HGNC:HGNC:19404]

This gene encodes a member of the NADPH oxidase family. The protein contains a PX domain and two SH3 domains. [provided by RefSeq, Jun 2012]

GeneCards Summary [HGNC:HGNC:19404]

NOXO1 (NADPH oxidase organizer 1) and Phagocytosis annotations removed.

UniProtKB/Swiss-Prot Summary [HGNC:HGNC:19404]

Constitutively expressed in all tissues. It is found in crystalline structures in the cytoplasmic membrane. Together with NOXO3 it is involved in generating reactive oxygen species (ROS). [provided by RefSeq, Jun 2012]

Gene Wiki entry [HGNC:HGNC:19404]

Additional gene information [HGNC:HGNC:19404]

Open Targets Platform [HGNC:HGNC:19404]

Alliance of Genotyped and Imputed Samples [HGNC:HGNC:19404]

Official Symbol NOXO1 provided by HGNC

Official Full Name NADPH oxidase organizer 1 provided by HGNC

Primary source HGNC:HGNC:19404

See related Ensembl:ENSG00000196408 MIM:611256; AllianceGenome:HGNC:19404

Gene type protein coding

RefSeq status REVIEWED

Organism Homo sapiens

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

Also known as SNX28; P41NOX; P41NOXA; P41NOXB; P41NOXC; SH3PXD5

Summary 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]

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

Orthologs mouse all

Look up genes in databases

The screenshot shows a gene database interface for the NOXO1 gene. The top navigation bar includes links for 'GENE', 'HOME', 'LOG IN', and 'HELP'. Below the navigation, the gene identifier 'noxo1' is displayed, followed by the gene name 'NOXO1' and its symbol 'NOXE1'. A blue header bar says 'Summaries for NOXO1 Gene'. On the left, a sidebar lists gene details: ID (noxo1), Name (noxo1), Symbol (NOXE1), Previous Name (none), Type (Synonym), Location (None), Description (Synonym), Genome Reference (None), Note (None), Comparative Information (None), and RGI (None). The main content area contains a bulleted list:

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

At the bottom, there are sections for 'Additional Information' (with a 'more' link) and 'Orthologs' (with links for 'mouse' and 'all').

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

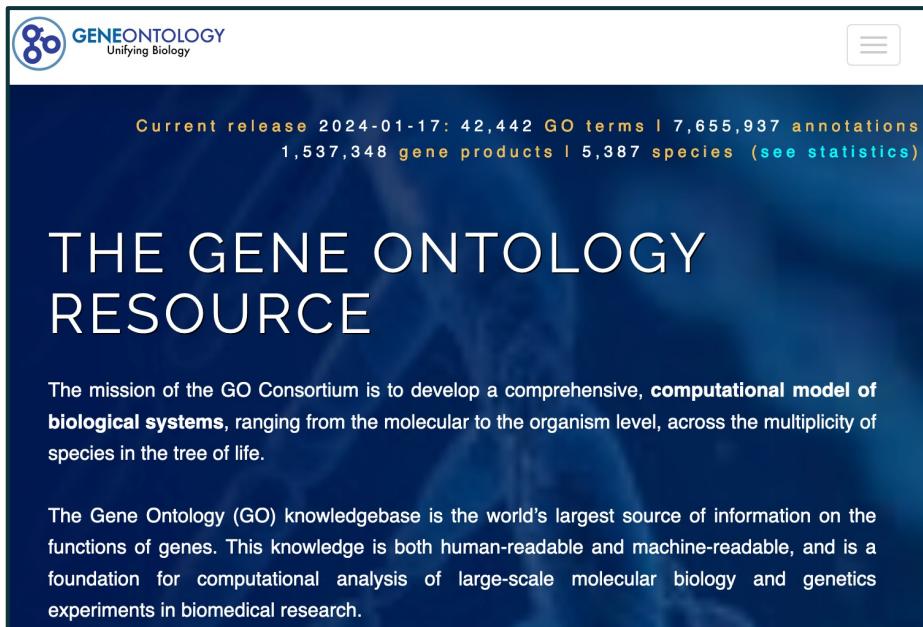
Functional gene sets

- 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
UniProtKB:A0A024QZG3	ATF5	involved_in	GO:0007623 circadian rhythm	ECO:0000265	GO_REF:0000107	Ensembl	BZIP domain-containing protein
UniProtKB:A0A024QZQ1	SIRT1	involved_in	GO:0007623 circadian rhythm	ECO:0000265	GO_REF:0000107	Ensembl	Deacetylase sirtuin-type domain-containing protein
UniProtKB:A0A024R230	NTRK2	involved_in	GO:0007623 circadian rhythm	ECO:0000265	GO_REF:0000107	Ensembl	Tyrosine-protein kinase receptor
UniProtKB:A0A024R241	NFIL3	involved_in	GO:0007623 circadian rhythm	ECO:0000256	GO_REF:0000002	InterPro	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 homepage of the Gene Ontology (GO) website. At the top left is the GO logo (three overlapping circles) and the text "GENEONTOLOGY Unifying Biology". In the top right corner is a menu icon. Below the header, a dark blue banner displays current statistics: "Current release 2024-01-17: 42,442 GO terms | 7,655,937 annotations | 1,537,348 gene products | 5,387 species" followed by a link "(see statistics)". The main title "THE GENE ONTOLOGY RESOURCE" is centered in large white letters. A descriptive paragraph below the title 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." Another paragraph at the bottom explains: "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

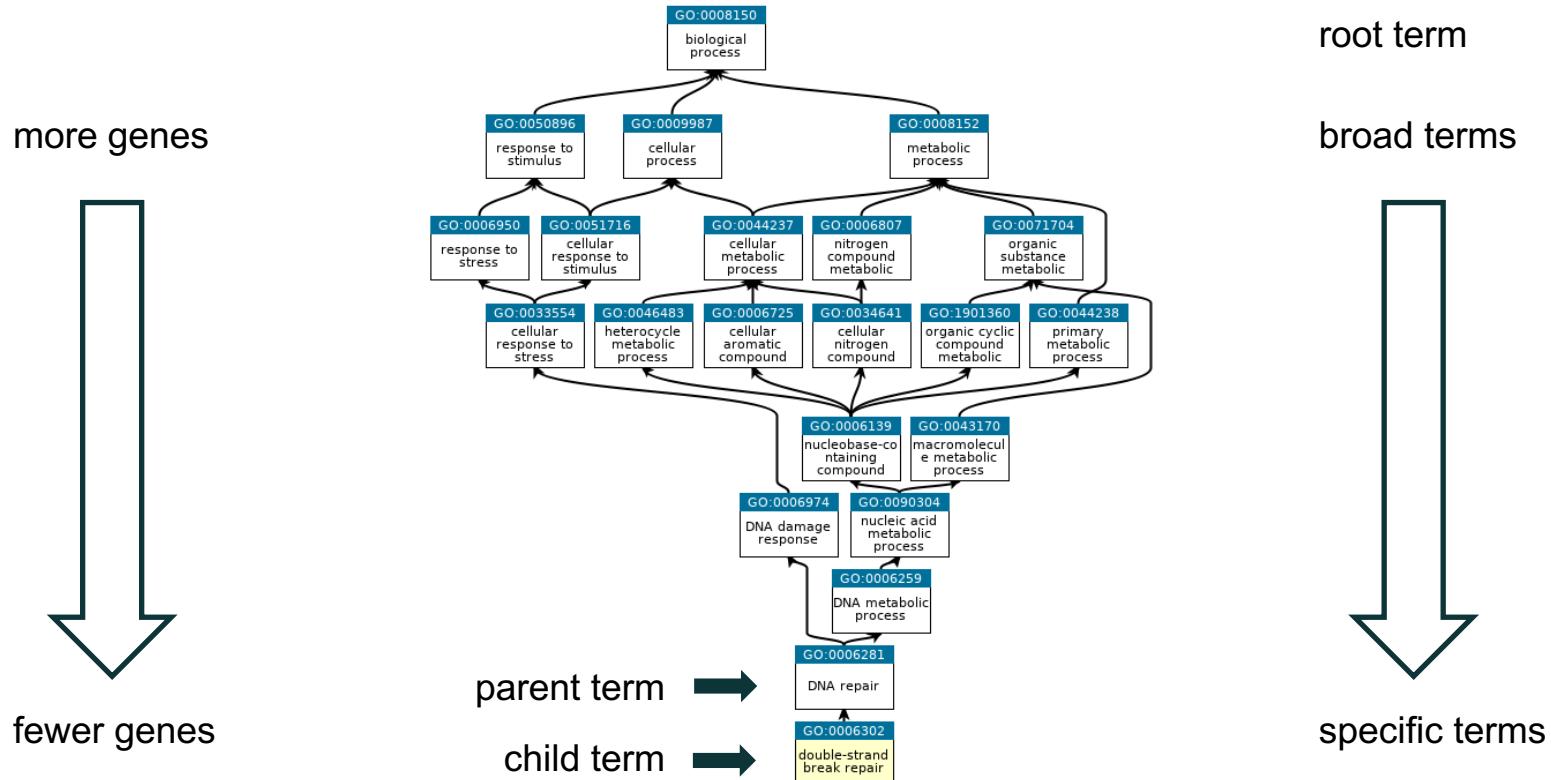
Current release 2024-01-17: 42,442 GO terms | 7,655,937 annotations
1,537,348 gene products | 5,387 species (see statistics)

- 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



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
GO:0002020	protease binding	IPI	UniProt
GO:0003677	DNA binding	IEA	UniProt
GO:0003697	single-stranded DNA binding	IDA	UniProt
GO:0005515	protein binding	IPI	IntAct
GO:0008022	protein C-terminus binding	IDA	MGI
GO:0010484	H3 histone acetyltransferase activity	IDA	UniProt
GO:0010485	H4 histone acetyltransferase activity	IDA	UniProt
GO:0042802	identical protein binding	IPI	IntAct
GO:0043015	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			
GO:0002020	protease binding	IPI	UniProt			
GO:0003677	DNA binding	IEA	UniProt			
GO:0003697	single-stranded DNA binding	IDA	UniProt			
GO:0005515	protein binding	IPI	IntAct			
GO:0008022	protein C-terminus binding	IDA	MGI			
GO:0010484	H3 histone acetyltransferase activity	IDA	UniProt			
GO:0010485	H4 histone acetyltransferase activity	IDA	UniProt			
GO:0042802	identical protein binding	IPI	IntAct			
GO:0043015	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	Term	Evidence	Annotation source
GO:0000152	nuclear ubiquitin ligase complex	IDA	ComplexPortal
GO:0000781	chromosome, telomeric region	IDA	BHF-UCL
GO:0000800	lateral element	IDA	MGI
GO:0005634	nucleus	IDA, IEA	UniProt
GO:0005654	nucleoplasm	IDA	HPA
GO:0005694	chromosome	IEA	Ensembl
GO:0005737	cytoplasm	IEA	UniProt
GO:0005813	centrosome	IDA	UniProt
GO:0005815	microtubule organizing center	IEA	UniProt
GO:0005829	cytosol	IDA	HPA
GO:0005856	cytoskeleton	IEA	UniProt
GO:0030141	secretory granule	IDA	UniProt
GO:0032991	protein-containing complex	IDA	MGI
GO:0033593	BRCA2-MAGE-D1 complex	IDA	UniProt
GO:1990391	DNA repair complex	IPI	ComplexPortal

BRCA2 example

Gene: BRCA2 ENSG00000139618

Description

BRCA2 DNA
Symbol;Acc:

Gene Synonyms

BRCC2, FAC

Location

Chromosome
GRCh38:CM

About this gene

This gene has
orthologues a

Transcripts

Show tran

GO: Molecular function ?

Show/hide columns (3 hidden)

Accession Term

GO:0002020	protease binding
GO:0003677	DNA binding
GO:0003697	single-stranded DNA binding
GO:0005515	protein binding
GO:0008022	protein C-terminus binding
GO:0010484	H3 histone acetyltransferase activity
GO:0010485	H4 histone acetyltransferase activity
GO:0042802	identical protein binding
GO:0043015	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
GO:0000722	telomere maintenance via recombination	IEA	Ensembl
GO:0000724	double-strand break repair via homologous recombination	IEA	
GO:0001556	oocyte maturation	IEA	Ensembl
GO:0001833	inner cell mass cell proliferation	IEA	Ensembl
GO:0006281	DNA repair	IEA	
GO:0006289	nucleotide-excision repair	IMP	UniProt
GO:0006302	double-strand break repair	IMP	UniProt
GO:0006310	DNA recombination	IEA	UniProt
GO:0006355	regulation of DNA-templated transcription	IBA	GO_Central
GO:0006974	cellular response to DNA damage stimulus	IEA	UniProt
GO:0006978	DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator	IEA	Ensembl
GO:0007049	cell cycle	IEA	UniProt

00000139618

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

Component ?

Term	Evidence	Annotation source
bar ubiquitin ligase complex	IDA	ComplexPortal
telomeric region	IDA	BHF-UCL
element	IDA	MGI
plasm	IDA, IEA	UniProt
osome	IEA	Ensembl
asm	IEA	UniProt
osome	IDA	UniProt
tube organizing center	IEA	UniProt
l	IDA	HPA
keleton	IEA	UniProt
ory granule	IDA	UniProt
n-containing complex	IDA	MGI
2-MAGE-D1 complex	IDA	UniProt
repair complex	IPI	ComplexPortal

Functional enrichment analysis

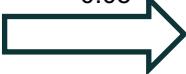
- How do we use all the existing annotation to interpret our gene list?
- Want to identify biological functions that are enriched in our gene list

Testing for functional enrichment

20,000 genes assayed

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG00000007769	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	3.686566857880831e-11
ENSDARG000000058731	7.374599119303466e-11
ENSDARG000000030896	1.069085837043268e-10
ENSDARG000000117300	3.836221667423099e-10
ENSDARG000000102758	7.675777258160306e-10
ENSDARG00000004754	1.3907777984676073e-9
ENSDARG000000099960	2.0021043563850804e-9
ENSDARG000000056615	2.032130501204873e-9
ENSDARG000000073820	3.749723988428957e-9
ENSDARG00000003570	4.02761509700461e-9
ENSDARG000000077799	7.49209763602349e-9
ENSDARG000000053836	7.754409809629005e-9
ENSDARG000000094678	7.81554132440011e-9
ENSDARG000000076914	2.8102379001185536e-8
ENSDARG000000037421	2.99098513534242e-8
ENSDARG000000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG000000018491	4.275633454932327e-8
ENSDARG000000109648	4.9384742488335793e-8
ENSDARG000000010231	5.822859503920341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.2504789486538527e-7
ENSDARG000000054196	1.356347303797386e-7
ENSDARG000000090548	1.8758159693858449e-7
ENSDARG000000104919	2.2367046789114162e-7
ENSDARG000000062788	3.7107672341197336e-7
ENSDARG000000079227	4.3287909103165297e-7
ENSDARG000000051888	4.6980261652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG000000056196	4.7089485706046475e-7
ENSDARG000000018283	4.917667288531457e-7
ENSDARG000000105731	5.285812733974555e-7
ENSDARG00000004954	5.526978490118169e-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	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG00000007769	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	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000040480	5.735034663692255e-14
ENSDARG00000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG000000013670	2.318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11
ENSDARG0000000004748	3.38076645838105e-11
ENSDARG000000102808	3.652027442583773e-11
ENSDARG000000059294	3.686666857880831e-11
ENSDARG000000058731	7.37459119303466e-11
ENSDARG000000030896	1.069085837043268e-10
ENSDARG0000000117300	3.836221667423099e-10
ENSDARG000000102758	7.675777258160306e-10
ENSDARG00000004754	1.3907777984676073e-9
ENSDARG000000099960	2.0021043563850804e-9
ENSDARG000000056615	2.03213051204873e-9
ENSDARG000000073820	3.749723988428957e-9
ENSDARG00000003570	4.02761509700461e-9
ENSDARG000000077799	7.49209763602349e-9
ENSDARG000000053836	7.754409809629005e-9
ENSDARG000000094678	7.81554132440011e-9
ENSDARG000000076914	2.8102379001185536e-8
ENSDARG000000037421	2.99098513534242e-8
ENSDARG000000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG0000000018491	4.27563345493327e-8
ENSDARG000000109648	4.9384742488335793e-8
ENSDARG000000010231	5.822859503920341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.2504789486538527e-7
ENSDARG000000054196	3.36547303797386e-7
ENSDARG000000090548	4.1758159693854849e-7
ENSDARG000000104919	2.2367046789114162e-7
ENSDARG000000062788	3.7107672341197336e-7
ENSDARG000000079227	4.3287909103165297e-7
ENSDARG000000051888	4.69802161652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG000000056196	4.7089485706046475e-7
ENSDARG000000018283	4.91766728531457e-7
ENSDARG000000105731	5.2858127233974355e-7
ENSDARG00000004954	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	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG00000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG000000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11

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

→ 2000/20000 = 10%

(18,000 not annotated to DNA repair)

200 genes annotated to DNA repair

200/500 = 40%

(300 not annotated to DNA repair)

Testing for functional enrichment

20,000 genes assayed

Gene	adjp
ENSDARG000000041294	1.0867269119101765e-32
ENSDARG000000060498	1.2517176061993635e-21
ENSDARG000000031683	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000040480	5.735034663692255e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG00000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG000000013670	2.3318547986985375e-11
ENSDARG000000039490	3.047413661053603e-11
ENSDARG00000002004940	3.047413661053603e-11
ENSDARG000000004748	3.38076645838105e-11
ENSDARG000000102808	3.652027442583773e-11
ENSDARG000000059294	3.686566857880831e-11
ENSDARG000000058731	7.374599119303466e-11
ENSDARG000000030896	1.069085837043268e-10
ENSDARG000000117300	3.836221667423099e-10
ENSDARG000000102758	7.675777258160306e-10
ENSDARG00000004754	1.3907777984676073e-9
ENSDARG000000099960	2.0021043563850804e-9
ENSDARG000000056615	2.03213051204873e-9
ENSDARG000000073820	3.749723988428957e-9
ENSDARG00000003570	4.027615069700461e-9
ENSDARG000000077799	4.9209763602349e-9
ENSDARG000000053836	7.75409809629005e-9
ENSDARG000000094678	7.8155413244001e-9
ENSDARG000000076914	8.102379001185536e-8
ENSDARG000000037421	2.99098513534242e-8
ENSDARG000000014340	3.505316564907291e-8
ENSDARG000000104773	4.1069391754303735e-8
ENSDARG000000105749	4.1069391754303735e-8
ENSDARG000000018491	4.27563345493327e-8
ENSDARG000000109648	4.9384742488335793e-8
ENSDARG000000010231	5.822859503920341e-8
ENSDARG000000039142	7.665381608460826e-8
ENSDARG000000104672	1.250478948538276e-7
ENSDARG000000054196	1.36347303797386e-7
ENSDARG000000090548	1.8758159693858449e-7
ENSDARG000000104919	2.3670467891114162e-7
ENSDARG000000062788	3.7107672341197336e-7
ENSDARG000000079227	4.027615069700461e-7
ENSDARG000000051888	4.6980261652096274e-7
ENSDARG000000077169	4.7089485706046475e-7
ENSDARG000000056196	4.7089485706046475e-7
ENSDARG000000018283	4.91766728531475e-7
ENSDARG000000105731	5.285812733974355e-7
ENSDARG00000004954	5.26970490118169e-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	2.364463411626359e-21
ENSDARG000000077982	2.9548243274497384e-14
ENSDARG000000070480	5.735034663692255e-14
ENSDARG00000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101482	2.6986616800262944e-13
ENSDARG000000034503	5.567338300152267e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG00000039490	3.047413661053603e-11

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

2000/20000 = 10%

(18,000 not annotated to DNA repair)

200 genes annotated to DNA repair

200/500 = 40%

(300 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	adj.p
ENSDARG0000041294	1.0867269119101765e-32
ENSDARG0000060498	1.2517176019936356e-21
ENSDARG0000031635	2.3646341166339e-21
ENSDARG0000077932	2.9548243724745738e-14
ENSDARG0000070448	3.7530546369255e-14
ENSDARG0000077699	1.567821386346665e-13
ENSDARG000002042435	1.907530227976848e-11
ENSDARG00000101482	2.698616806202944e-11
ENSDARG00000304593	5.567333803152267e-11
ENSDARG0000013670	3.318547896853575e-11
ENSDARG00000329490	3.047413661051603e-11
ENSDARG0000047478	3.380766458381055e-11
ENSDARG00000102898	3.6520274255377e-11
ENSDARG00000502934	3.86666685788681e-11
ENSDARG0000058731	3.745941913034661e-11
ENSDARG00000300896	1.0698065837043268e-10
ENSDARG00000117300	3.83622167423090e-10
ENSDARG000002007278	7.67577725816306e-10
ENSDARG0000044754	1.3907772846760793e-9
ENSDARG000002009996	2.002104356385084e-9
ENSDARG0000056615	2.032130512084878e-9
ENSDARG0000073820	3.749723988428957e-9
ENSDARG00000300357	4.027615069700461e-9
ENSDARG0000077799	4.792097636203496e-9
ENSDARG00000503836	7.54504896626030e-9
ENSDARG0000094678	7.15154312440011e-9
ENSDARG00000706914	2.8102379001155536e-9
ENSDARG0000037241	2.99098155353424e-8
ENSDARG000001004340	3.50531654907912e-8
ENSDARG0000104773	4.106931754363735e-8
ENSDARG00000200105749	4.106931754363735e-8
ENSDARG0000018491	4.275633459343227e-8
ENSDARG00000109648	4.9384742480335793e-8
ENSDARG00000100251	5.422859509023041e-8
ENSDARG0000039142	7.66538168646862e-8
ENSDARG000001044672	1.2504789485653527e-8
ENSDARG0000054196	1.3563473037973786e-7
ENSDARG00000200090548	1.875185963854849e-7
ENSDARG00000104919	2.3760748789114162e-7
ENSDARG00000026776	3.710767234119736e-7
ENSDARG00000079272	4.3287909131652597e-7
ENSDARG00000105588	4.68802165265274e-7
ENSDARG0000027119	4.7089485706044675e-7
ENSDARG0000051696	4.7089485706044675e-7
ENSDARG00000100823	4.91766728531457e-7
ENSDARG00000105731	5.28512739734357e-7
ENSDARG00000049593	5.26970940118169e-7
ENSDARG0000025903	7.245067298625657e-7
ENSDARG0000020045094	7.99947486153812e-7
ENSDARG00000100504	8.037125469772779e-7

Adjusted
p-value
< 0.05

500 significantly DE genes

Gene	adjp
ENSDARG00000041294	1.0867269119101765e-32
ENSDARG00000060498	1.2517176061993635e-21
ENSDARG00000031683	2.364463411626339e-21
ENSDARG00000077982	2.9548243274497384e-14
ENSDARG00000070480	5.735034663692255e-14
ENSDARG00000007769	1.5678213864306653e-13
ENSDARG000000102435	1.9075360227976884e-13
ENSDARG000000101842	2.6986616800262944e-13
ENSDARG00000034503	5.5673383001522676e-13
ENSDARG00000013670	2.3318547986985375e-11
ENSDARG00000039490	3.047413661053603e-11

200 genes annotated
to DNA repair

$$200/500 = 40\%$$

(300 not annotated to
DNA repair)

2000 genes annotated
function (e.g. DNA)

$$2000/20000 = 10\%$$

(18,000 not annotated
DNA repair)

2000 genes annotated
function (e.g. DNA)

$$\longrightarrow \quad 2000/20000 = 10\%$$

(18,000 not annotated
DNA repair)

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

Hypergeometric test

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

```
> 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

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
- 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)

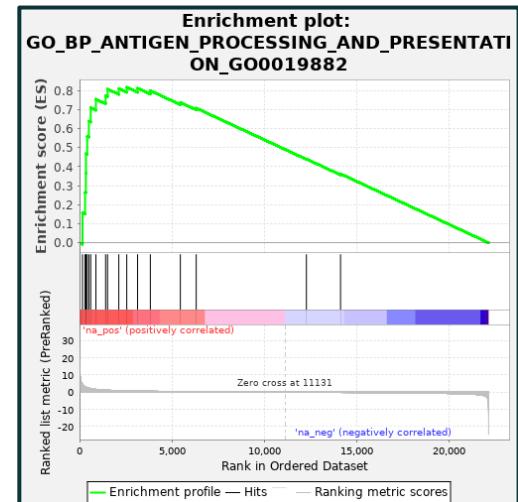
Functional enrichment tools

- 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 nearly 1000 species/strains



Other functional enrichment tools

- Other tools are available (and good):
 - Enrichr (maayanlab.cloud/Enrichr/):
 - Web-based
 - Similar to g:Profiler
 - Only human, mouse, fly, yeast, worm and zebrafish
 - GSEA (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
- Since September, g:Profiler had been using Ensembl 110, which came out in July last year
- But Ensembl 111 came out last month and g:Profiler was updated yesterday
- Recommend using Ensembl IDs as input, but not essential

g:Profiler

News Archives Beta API R client FAQ Docs Contact Cite g:Profiler Services using g:P GMT Helper

≡

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 [?](#)

Advanced options [▼](#)

Data sources [▼](#)

Bring your data (Custom GMT) [▼](#)

Run query random example mixed query example

g:Profiler – four tools

The screenshot shows the g:Profiler homepage with a red box highlighting the top navigation bar containing four tools: g:GOST, g:Convert, g:Orth, and g:SNPense.

g:GOST
Functional profiling

g:Convert
Gene ID conversion

g:Orth
Orthology search

g:SNPense
SNP id to gene name

Below the tools, there are input fields for "Query", "Upload query", and "Upload bed file". A note says "Input is whitespace-separated list of genes".

Options

Organism: Homo sapiens (Human)

Highlight driver terms in GO

Ordered query

Run as multiquery

Advanced options

Data sources

Bring your data (Custom GMT)

At the bottom, there are buttons for "Run query", "random example", and "mixed query example".

g:Profiler – gene list

The screenshot shows the g:Profiler interface for a gene list analysis. The top navigation bar includes links for News, Archives, Beta, API, R client, FAQ, Docs, Contact, Cite g:Profiler, Services using g:P, and GMT Helper. Below the navigation is a horizontal menu bar with four main options: g:GOSt (Functional profiling), g:Convert (Gene ID conversion), g:Orth (Orthology search), and g:SNPense (SNP id to gene name). The g:GOSt section is highlighted with an orange background. On the left, there are three input methods: Query (button), Upload query (link), and Upload bed file (link). A note below says "Input is whitespace-separated list of genes". A large red circle highlights this entire input area. To the right is the "Options" panel, which includes fields for Organism (set to Homo sapiens), Highlight (checkbox checked), Order by (checkbox unchecked), and Run as multiquery (checkbox unchecked). A callout box highlights the identifier recognition counts: "101 identifiers recognised for human", "80 for mouse; 92 for zebrafish". Below the options are buttons for Advanced options, Data sources, and Bring your data (Custom GMT).

g:Profiler

News Archives Beta API R client FAQ Docs Contact Cite g:Profiler Services using g:P GMT Helper

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

101 identifiers recognised for human

80 for mouse; 92 for zebrafish

Advanced options

Data sources

Bring your data (Custom GMT)

Run query random example mixed query example

g:Profiler – options

g:Profiler

News Archives Beta API R client FAQ Docs Contact Cite g:Profiler Services using g:P GMT Helper

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 ?

984 species/strains in current release

Options

Organism: ?

Homo sapiens (Human)

Highlight driver terms in GO ?

Ordered query ?

Run as multiquery ?

Advanced options ▾

Data sources ▾

Bring your data (Custom GMT) ▾

Run query random example mixed query example

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

Organism: ?

Homo sapiens (Human)

Highlight driver terms in GO ?

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 ▾

select all clear all 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

The screenshot shows the g:Profiler web application interface. At the top, there are three navigation buttons: "Query" (highlighted in green), "Upload query", and "Upload bed file". Below these is a text input field with placeholder text "Input is whitespace-separated list of genes ?". Underneath the input field are three buttons: "Run query" (orange), "random example", and "mixed query example". On the right side, there is a panel titled "Options" containing settings for the organism (set to "Homo sapiens (Human)"), highlighting driver terms in GO (checked), and other query options like "Ordered query" and "Run as multiquery". A "Data sources" section is also present. The main feature highlighted by a large red circle is the "Bring your data (Custom GMT)" section. This section contains a dashed-dotted box for dragging a GMT file or browsing for one, and fields for "File name used*" and "or insert token:" with a dropdown menu showing "not selected". A note at the bottom states: "* If changed query will be run across the selected GMT. For public GMT sources see link".

Query Upload query Upload bed file

Input is whitespace-separated list of genes ?

Run query random example mixed query example

Organism: Homo sapiens (Human)

Highlight driver terms in GO ?

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*

or insert token:

not selected

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

g:Profiler – documentation

g:Profiler News Archives Beta API R client **FAQ Docs** Contact Cite g:Profiler Services using g:P GMT Helper ≡

g:GOST
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:GOST**, 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:GOST
Using g:GOST
Highlighting
Examples
g:Convert
Using g:Convert
Examples
g:Orth
Using g:Orth
Examples
g:SNPense

g:Profiler – archives

The screenshot shows the g:Profiler website interface. At the top, there is a navigation bar with links: News, Archives (which is highlighted with a red box), Beta, API, R client, FAQ, Docs, Contact, Cite g:Profiler, Services using g:P, GMT Helper, and a three-line menu icon. Below the navigation bar is a row of four buttons: g:GOSt (Functional profiling), g:Convert (Gene ID conversion), g:Orth (Orthology search), and g:SNPense (SNP id to gene name). The main content area has a dark header with the word "Archives". Below the header, there is a paragraph of text followed by a bulleted list of Ensembl versions.

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 [110](#), Ensembl Genomes [57](#) (database built on 2023-09-14)
- Ensembl [109](#), Ensembl Genomes [56](#) (database built on 2023-03-29)
- Ensembl [108](#), Ensembl Genomes [55](#) (database built on 2022-12-28)
- Ensembl [107](#), Ensembl Genomes [54](#) (database built on 2022-09-15)
- 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)

g:Profiler – API and libraries

The screenshot shows the g:Profiler website interface. At the top, there is a navigation bar with links: News, Archives, Beta, API (which is highlighted with a red box), R client, FAQ, Docs, Contact, Cite g:Profiler, Services using g:P, GMT Helper, and a menu icon. Below the navigation bar, there are four main service buttons: g:GOST (Functional profiling), g:Convert (Gene ID conversion), g:Orth (Orthology search), and g:SNPense (SNP id to gene name). The main content area has a title "g:Profiler client libraries". It contains three sections: "R client", "Python client", and "g:Profiler API". The "R client" section describes the gprofiler2 library and provides a link to the help page. The "Python client" section describes the gprofiler-official library and provides a link to the package description. The "g:Profiler API" section states that requests are generally made as POST requests with a JSON body and return JSON output. To the right of the main content, there is a sidebar titled "Contents" which lists various API endpoints with their corresponding examples and descriptions.

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](#)

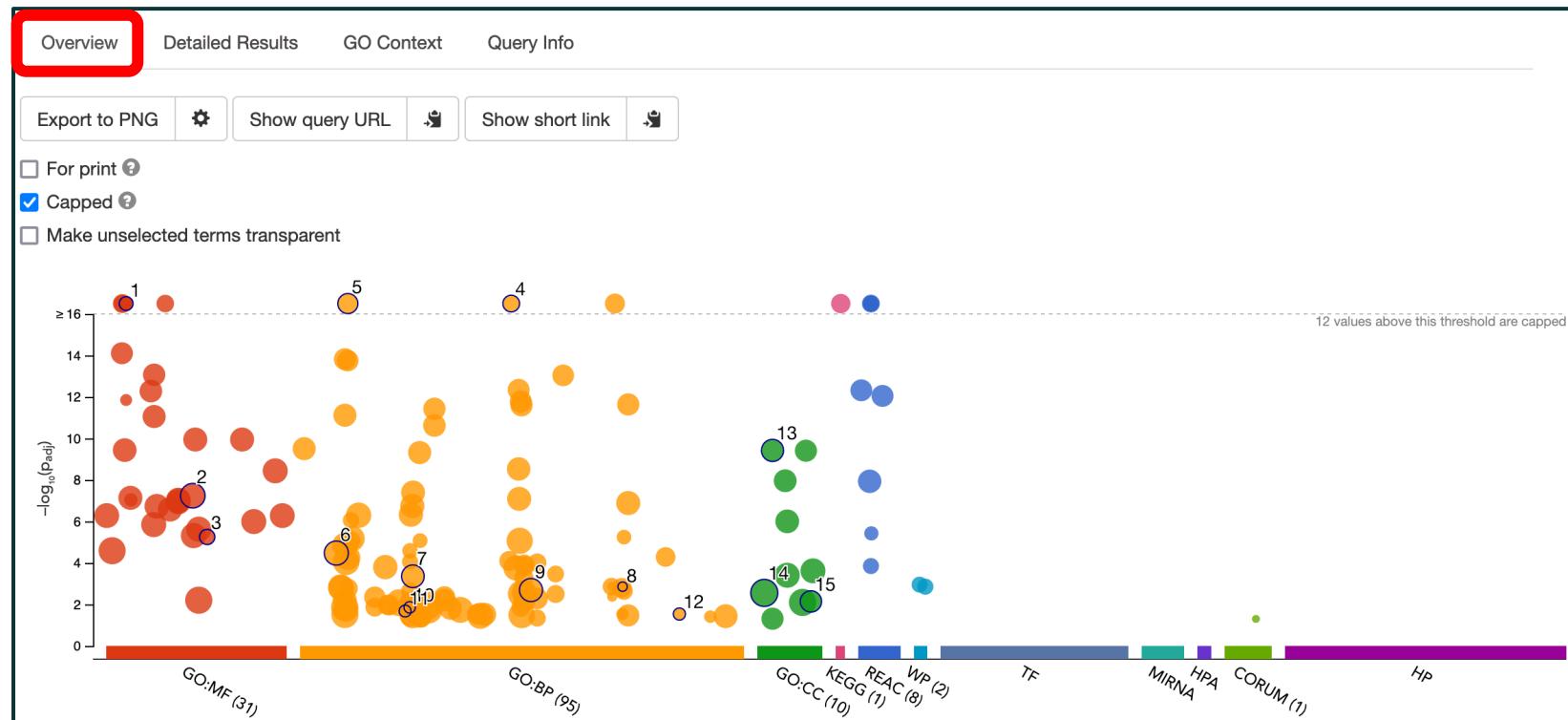
g:Profiler API

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

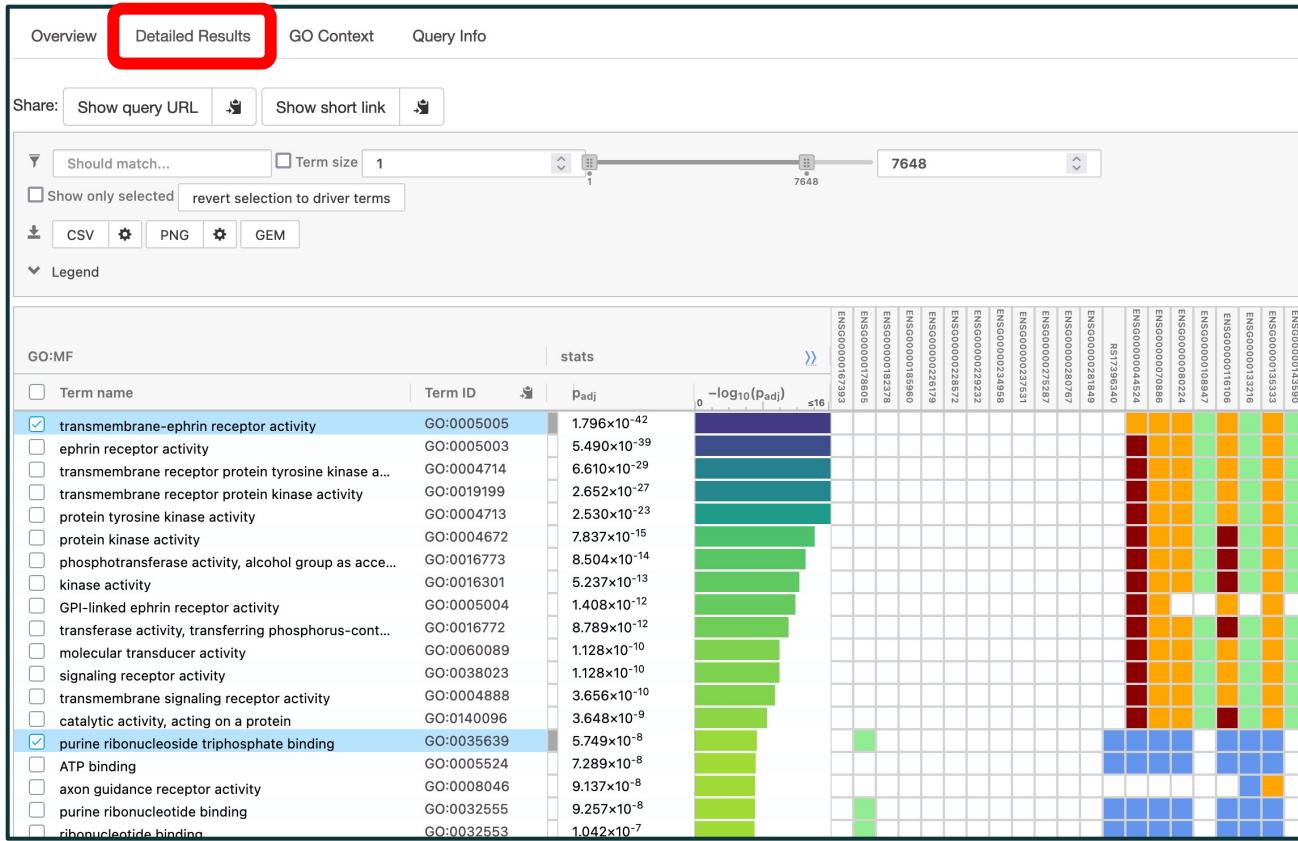
Contents

- R client
 - 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 – overview



g:Profiler – detailed results



g:Profiler - GO context

Overview Detailed Results **GO Context** Query Info

GO:MF > (1 2 3 / 3) Tip: hover a node to display its detailed information here

```
graph TD; GO-0008041[GO-0008041 ionotropic receptor] --> GO-0005241[GO-0005241 ATPase activity]; GO-0008041 --> GO-0008042[GO-0008042 G-protein coupled receptor]; GO-0005241 --> GO-0018742[GO-0018742 hydrolase activity, hydrolyzing ester bonds]; GO-0005241 --> GO-0018772[GO-0018772 hydrolase activity, hydrolyzing ester bonds, EC 3.1.1.12]; GO-0018742 --> GO-0016501[GO-0016501 ester hydrolase, EC 3.1.1.13]; GO-0018742 --> GO-0016512[GO-0016512 ester hydrolase, EC 3.1.1.14]; GO-0018772 --> GO-0009099[GO-0009099 esterase hydrolase, EC 3.1.1.14]; GO-0018772 --> GO-0018773[GO-0018773 esterase hydrolase, EC 3.1.1.14]; GO-0016501 --> GO-0009092[GO-0009092 esterase, EC 3.1.1.15]; GO-0016512 --> GO-0009093[GO-0009093 esterase, EC 3.1.1.15]; GO-0018773 --> GO-0009094[GO-0009094 esterase, EC 3.1.1.15]; GO-0018773 --> GO-0011199[GO-0011199 esterase, EC 3.1.1.15]; GO-0009092 --> GO-0009074[GO-0009074 esterase inhibitor, EC 3.1.1.15]; GO-0009093 --> GO-0009084[GO-0009084 esterase inhibitor, EC 3.1.1.15]; GO-0009094 --> GO-0009085[GO-0009085 esterase inhibitor, EC 3.1.1.15]; GO-0011199 --> GO-0009083[GO-0009083 esterase inhibitor, EC 3.1.1.15]; GO-0009083 --> GO-0009061[GO-0009061 esterase inhibitor, EC 3.1.1.15]; GO-0009084 --> GO-0009061[GO-0009061 esterase inhibitor, EC 3.1.1.15];
```

g:Profiler - beta

The screenshot shows the g:Profiler beta website. The top navigation bar includes links for News, Archives, Stable (highlighted with a red box), API, R client, FAQ, Docs, Contact, Cite g:Profiler, Services using g:P, GMT Helper, and a menu icon. Below the navigation is a row of four buttons: g:GOSt (Functional profiling), g:Convert (Gene ID conversion), g:Orth (Orthology search), and g:SNPense (SNP id to gene name). The main content area features a large input field for genes, query options (Query, Upload query, Upload bed file), and an 'Options' section. The 'Options' section contains settings for organism (Homo sapiens), highlighting driver terms in GO, and running as a multiquery. It also includes three expandable sections: Advanced options, Data sources, and Bring your data (Custom GMT). At the bottom are buttons for Run query, random example, and mixed query example.

g:Profiler ^β

News Archives Stable API R client FAQ Docs Contact Cite g:Profiler Services using g:P GMT Helper ≡

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 ?

Organism: ?

Homo sapiens (Human)

Highlight driver terms in GO ?

Ordered query ?

Run as multiquery ?

Advanced options ▾

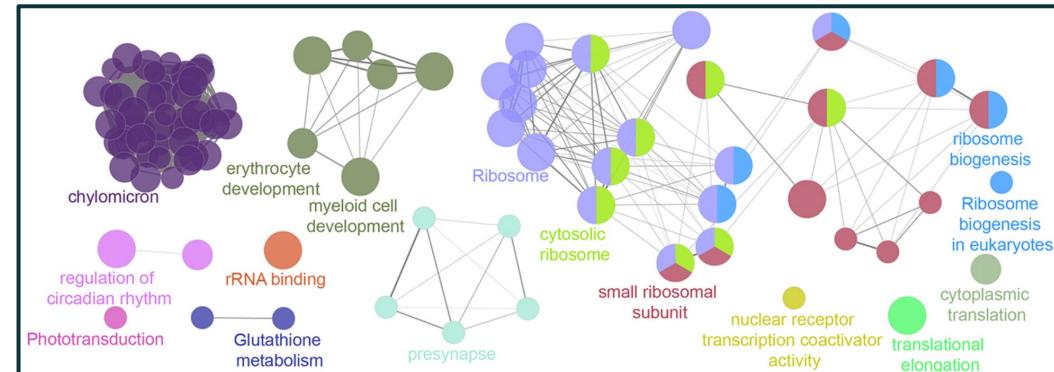
Data sources ▾

Bring your data (Custom GMT) ▾

Run query random example mixed query example

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
 - Revigo: <http://revigo.irb.hr/>



g:Profiler live demo!

- biit.cs.ut.ee/gprofiler/

Exercises (plus data and slides)

- Exercises are available from:

rnaseq2024.buschlab.org

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