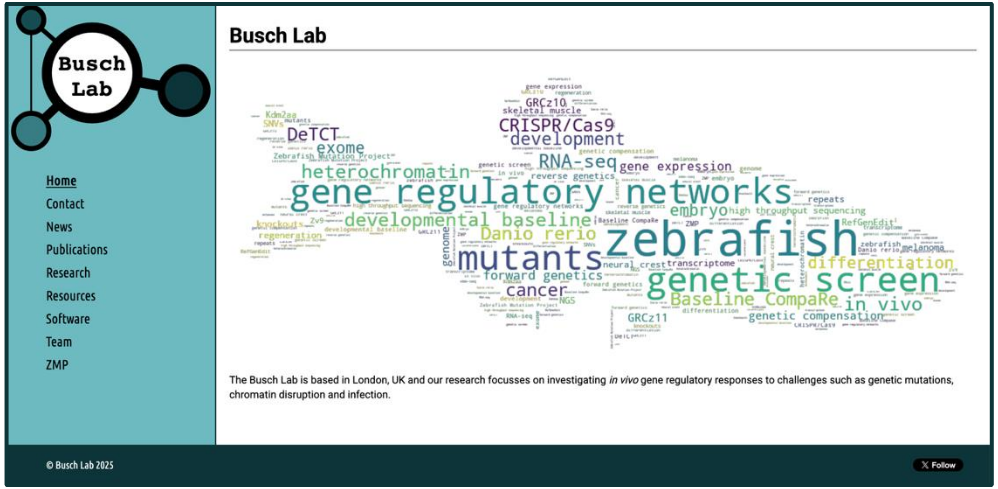
Introduction to RNA-seq and functional interpretation:

Next steps in gene prioritisation

12th Feb 2026

Me

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



Busch Lab

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.

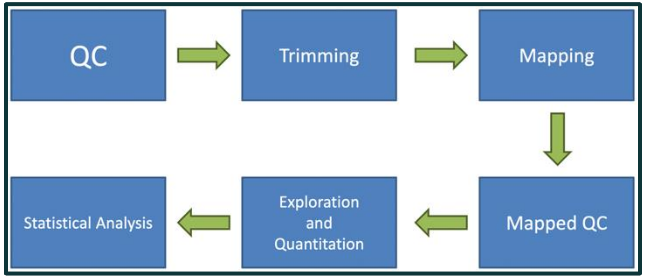
© Busch Lab 2025

Questions

- For timely questions, just unmute and ask
- Or add your question to the Q&A document and I’ll answer in a break or later

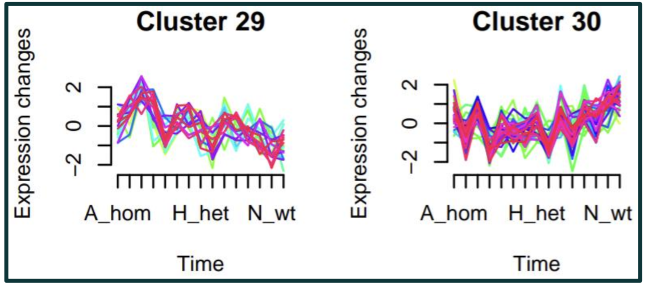
Gene list of interest

- 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…



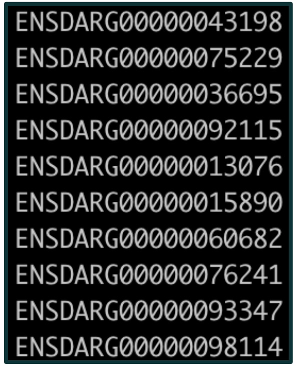
QC → Trimming → Mapping → Mapped QC → Exploration and Quantitation → Statistical Analysis

From: Simon Andrews

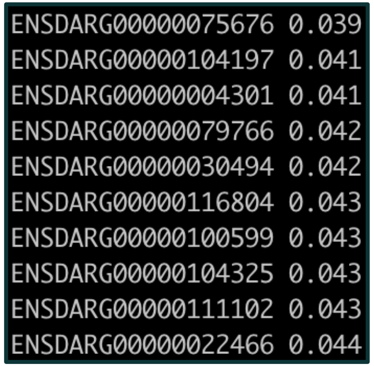


Unranked or ranked gene list?

- 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



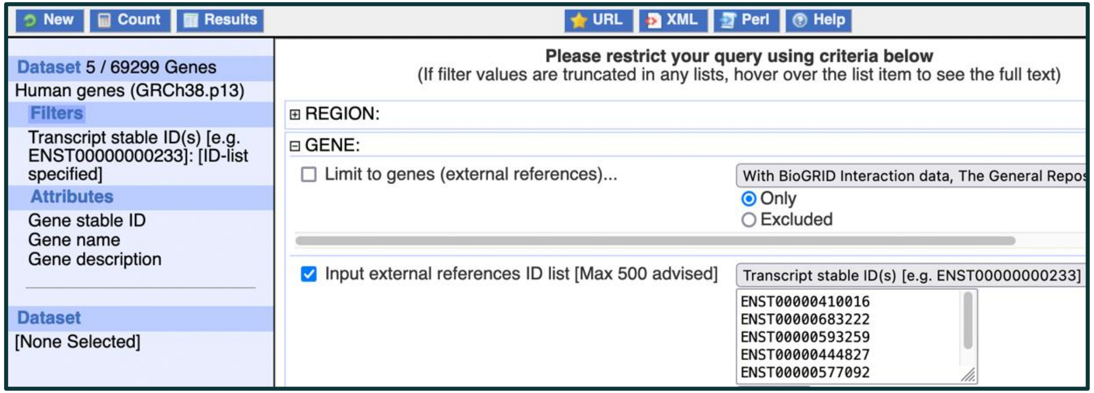
```
ENSDARG00000043198
ENSDARG00000075229
ENSDARG00000036695
ENSDARG00000092115
ENSDARG00000013076
ENSDARG00000015890
ENSDARG00000060682
ENSDARG00000076241
ENSDARG00000093347
ENSDARG00000098114
```



```
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



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:

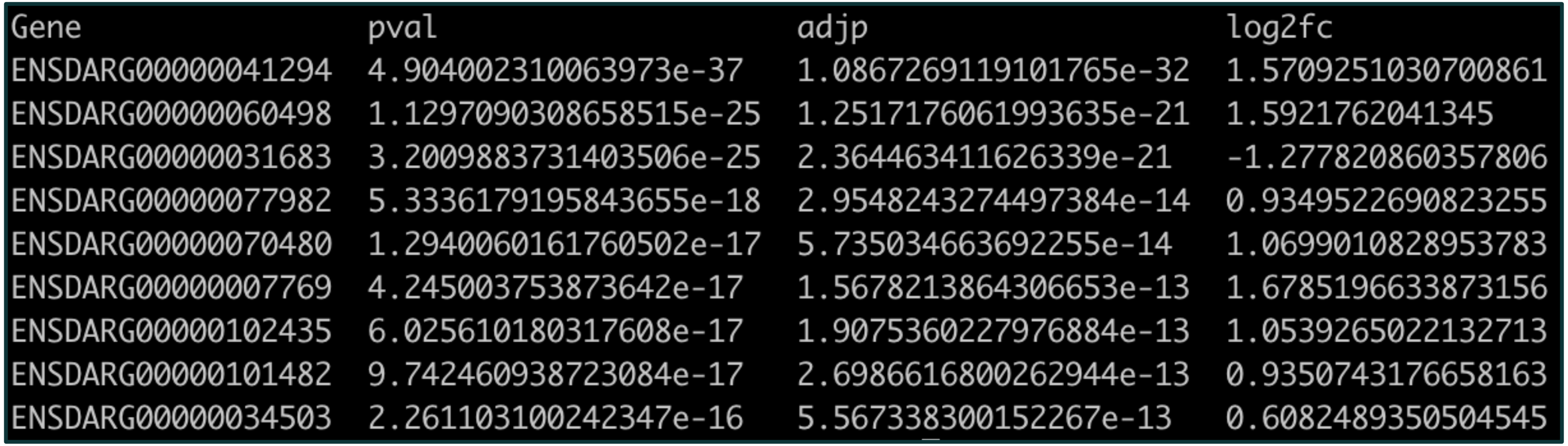
□ Limit to genes (external references)... With BioGRID Interaction data, The General Repos
● Only
○ Excluded

☑ Input external references ID list [Max 500 advised] Transcript stable ID(s) [e.g. ENST00000000233]

```
ENST00000410016
ENST00000683222
ENST00000593259
ENST00000444827
ENST00000577092
```

What next?

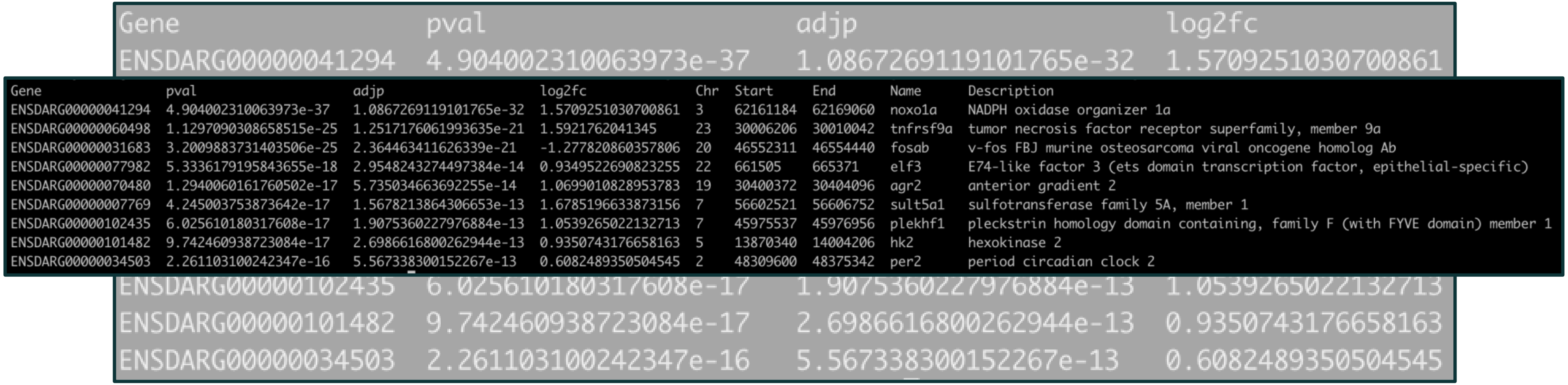
- Have a gene list, but what do you do next?
- How do you relate the gene list to existing knowledge?



```
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  2.36446341162633 9e-21  -1.277820860357806
ENSDARG00000077982  5.3336179195843655e-18  2.9548243274497384e-14  0.9349522690823255
ENSDARG00000070480  1.2940060161760502e-17  5.73503466369225 5e-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)

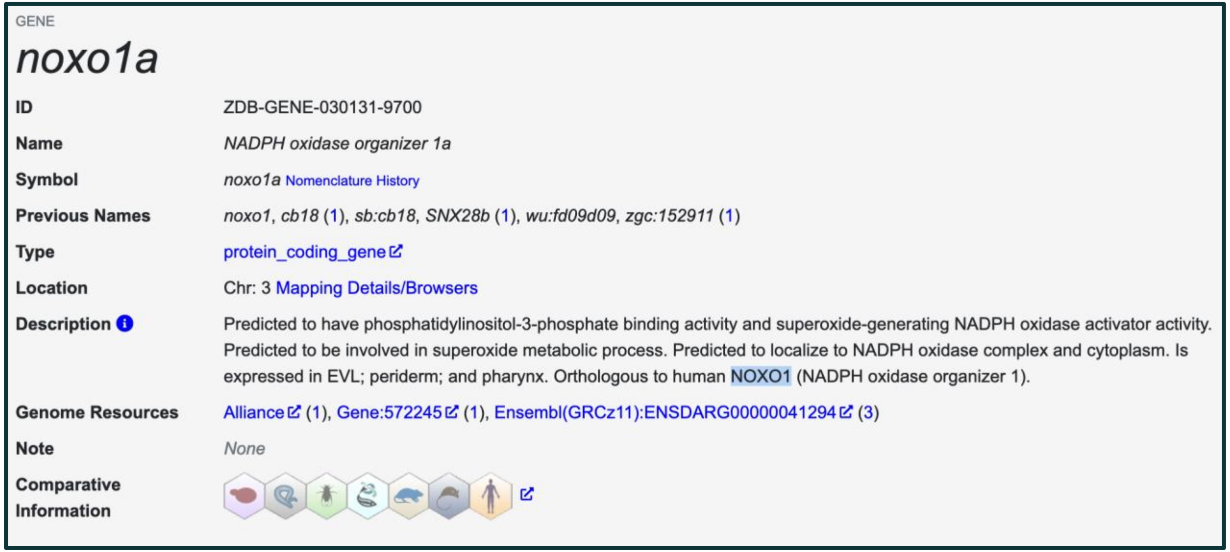


```
Gene                pval                   adjp                   log2fc
ENSDARG00000041294  4.904002310063973e-37  1.0867269119101765e-32 1.570925103070086l
```

```
Gene                pval                   adjp                   log2fc              Chr  Start     End       Name      Description
ENSDARG00000041294  4.904002310063973e-37  1.0867269119101765e-32 1.570925103070086l  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

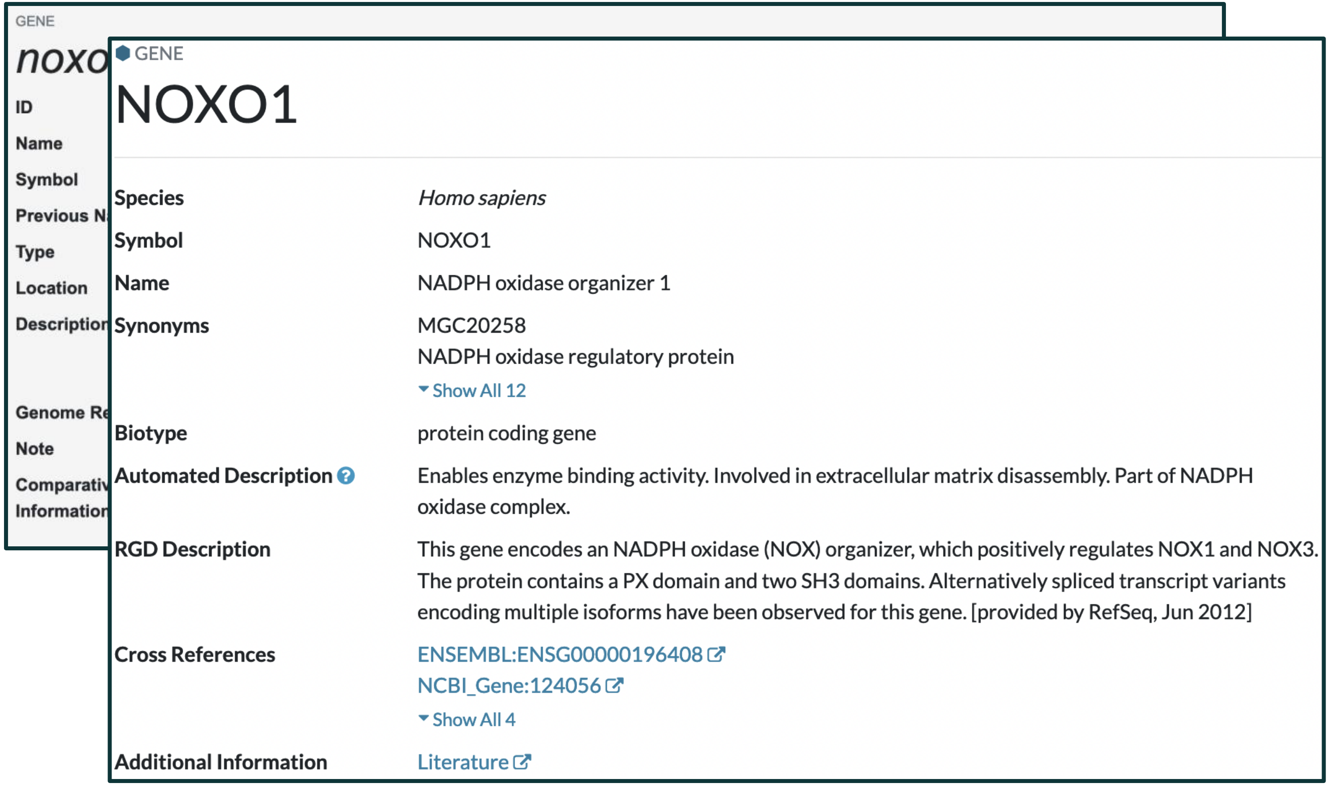


GENE

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

ID

Name

Symbol

Previous N

Type

Location

Description

Genome Re

Note

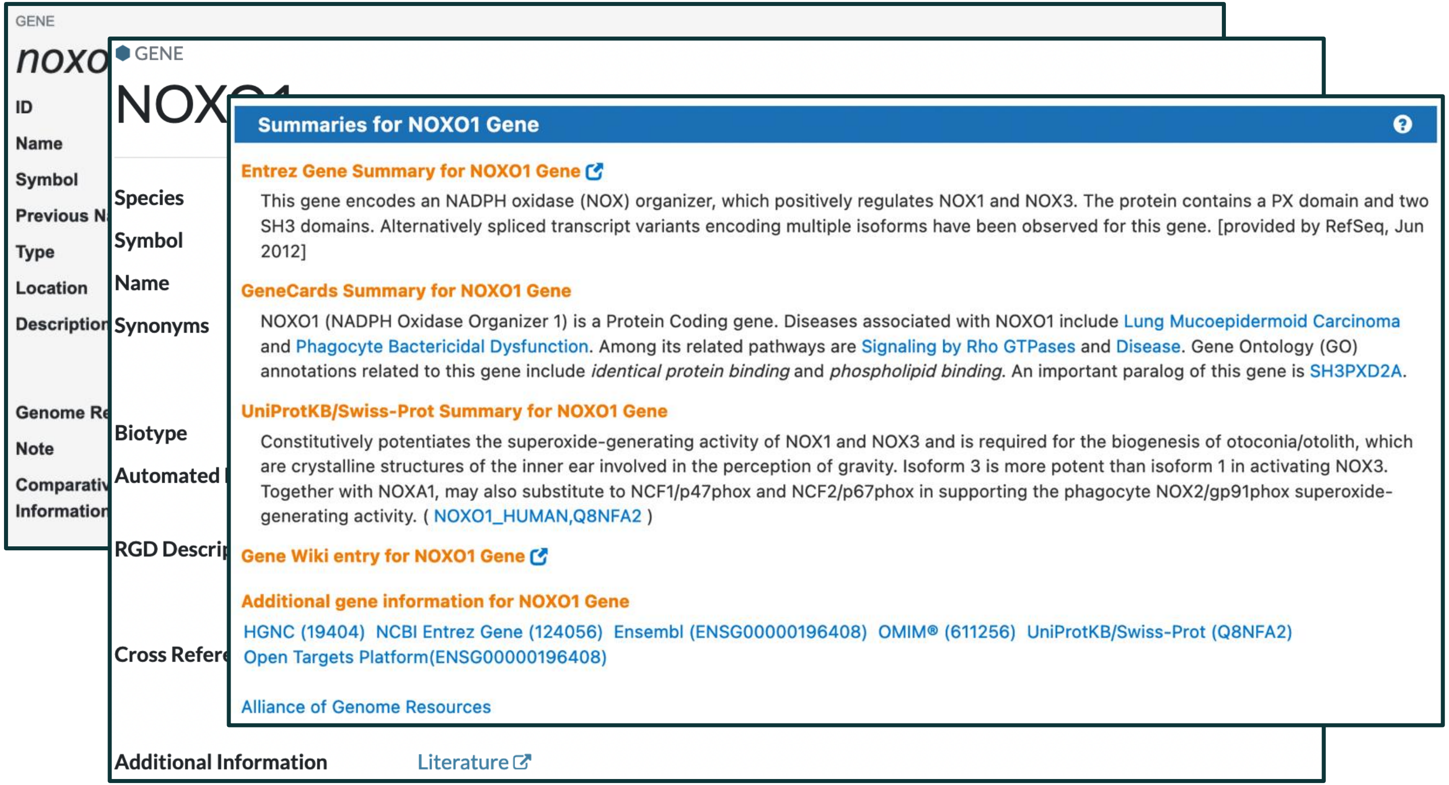
Comparativ Information

GENE

NOXO1

Species	*Homo sapiens*
Symbol	NOXO1
Name	NADPH oxidase organizer 1
Synonyms	MGC20258 NADPH oxidase regulatory protein ▾ Show All 12
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	ENSEMBL:ENSG00000196408 NCBI_Gene:124056 ▾ Show All 4
Additional Information	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 Refere

Additional Information Literature

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)

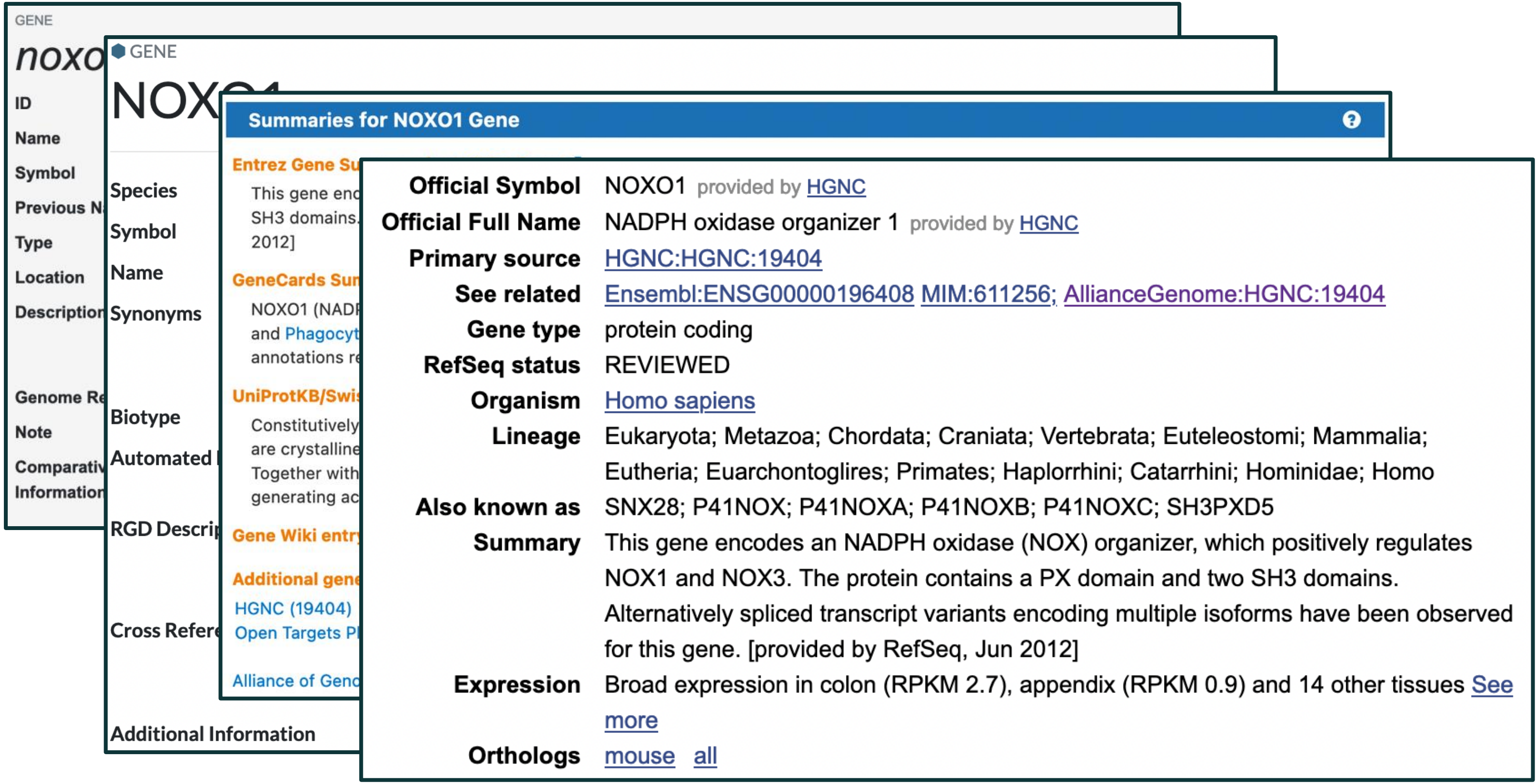
Gene Wiki entry for NOXO1 Gene

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

ID

Name

Symbol

Previous N

Type

Location

Description

Genome Re

Note

Comparativ Information

GENE

NOXO1

Species

Symbol

Name

Synonyms

Biotype

Automated

RGD Descrip

Cross Refere

Additional Information

Summaries for NOXO1 Gene

Entrez Gene Su

This gene enc SH3 domains. 2012]

GeneCards Sur

NOXO1 (NADP and Phagocyt annotations re

UniProtKB/Swis

Constitutively are crystalline Together with generating ac

Gene Wiki entr

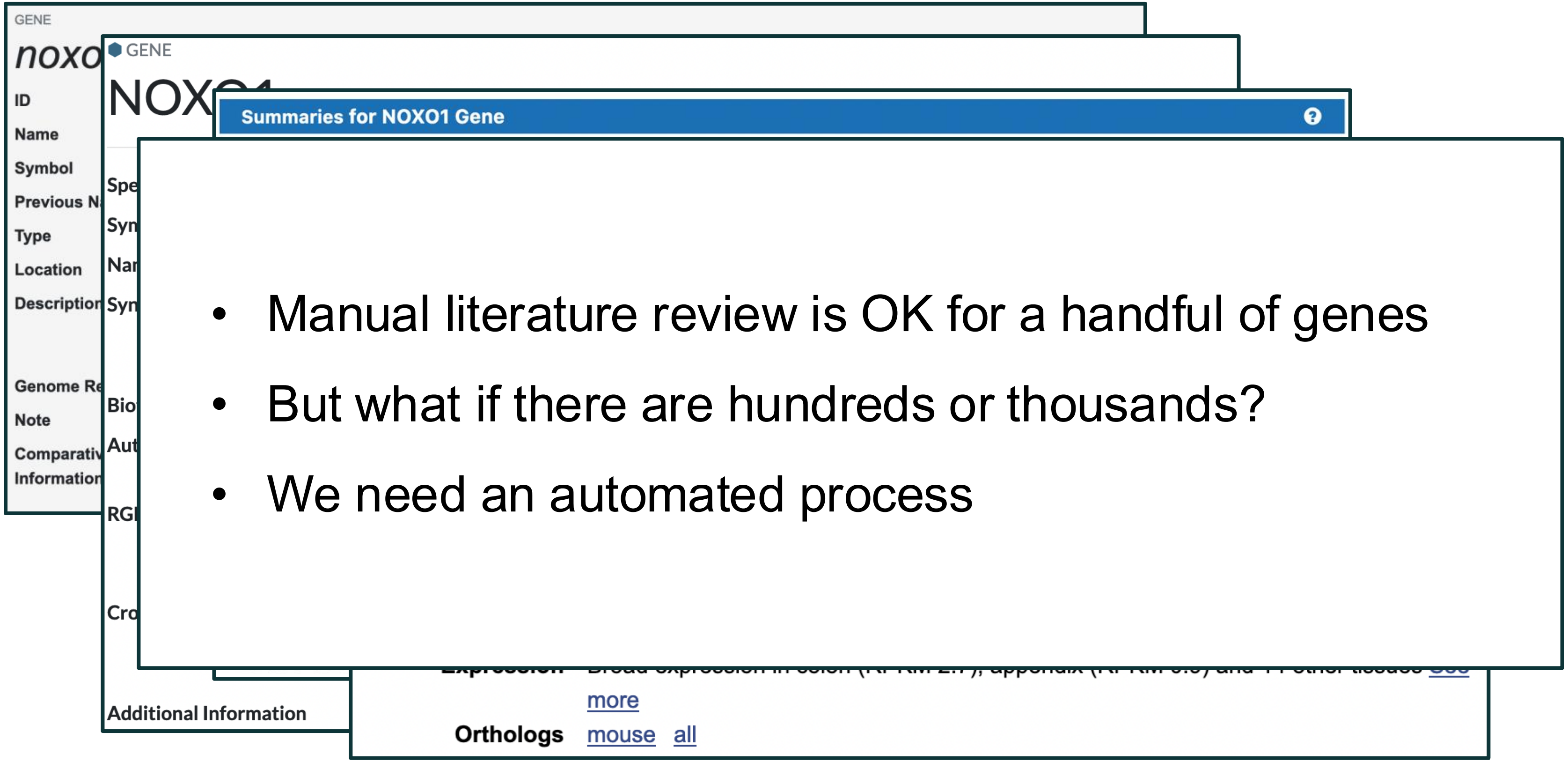
Additional gene

HGNC (19404)
Open Targets P

Alliance of Geno

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



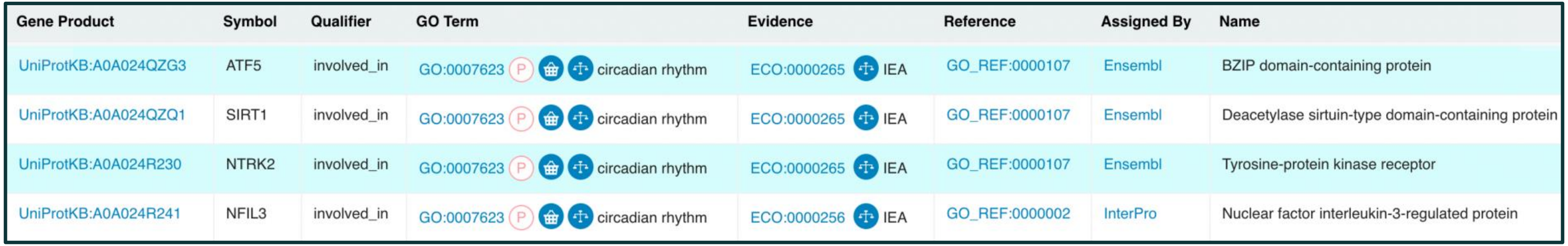
- Manual literature review is OK for a handful of genes
- But what if there are hundreds or thousands?
- We need an automated process

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

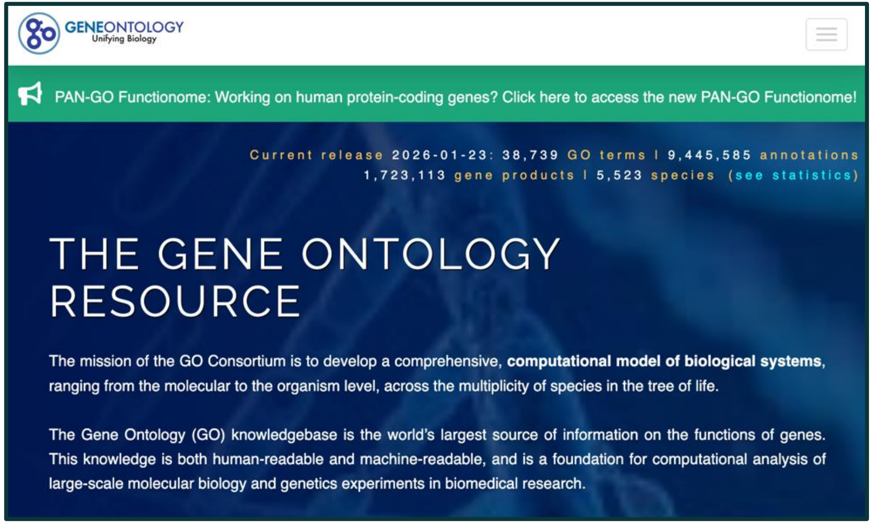
- Existing knowledge is organised into **functional gene sets** in a standardised way, using data from previous experiments
- A functional gene set is a group of genes with a common biological relationship (e.g. annotated to same biological process or involved in same pathway)
- e.g. circadian rhythm:



Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	Assigned By	Name
UniProtKB:A0A024QZG3	ATF5	involved_in	GO:0007623 Ⓟ circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	BZIP domain-containing protein
UniProtKB:A0A024QZQ1	SIRT1	involved_in	GO:0007623 Ⓟ circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	Deacetylase sirtuin-type domain-containing protein
UniProtKB:A0A024R230	NTRK2	involved_in	GO:0007623 Ⓟ circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	Tyrosine-protein kinase receptor
UniProtKB:A0A024R241	NFIL3	involved_in	GO:0007623 Ⓟ circadian rhythm	ECO:0000256 IEA	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



PAN-GO Functionome: Working on human protein-coding genes? Click here to access the new PAN-GO Functionome!

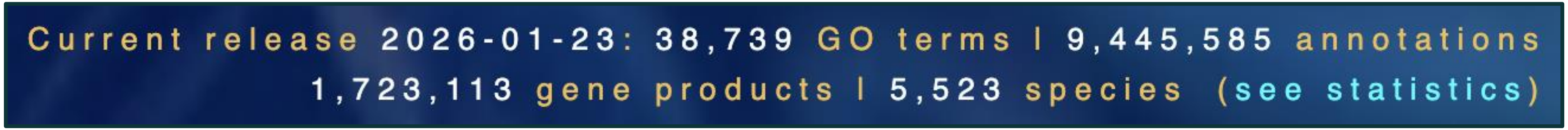
Current release 2026-01-23: 38,739 GO terms | 9,445,585 annotations
1,723,113 gene products | 5,523 species (see statistics)

THE GENE ONTOLOGY RESOURCE

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.

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 2026-01-23: 38,739 GO terms | 9,445,585 annotations
1,723,113 gene products | 5,523 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

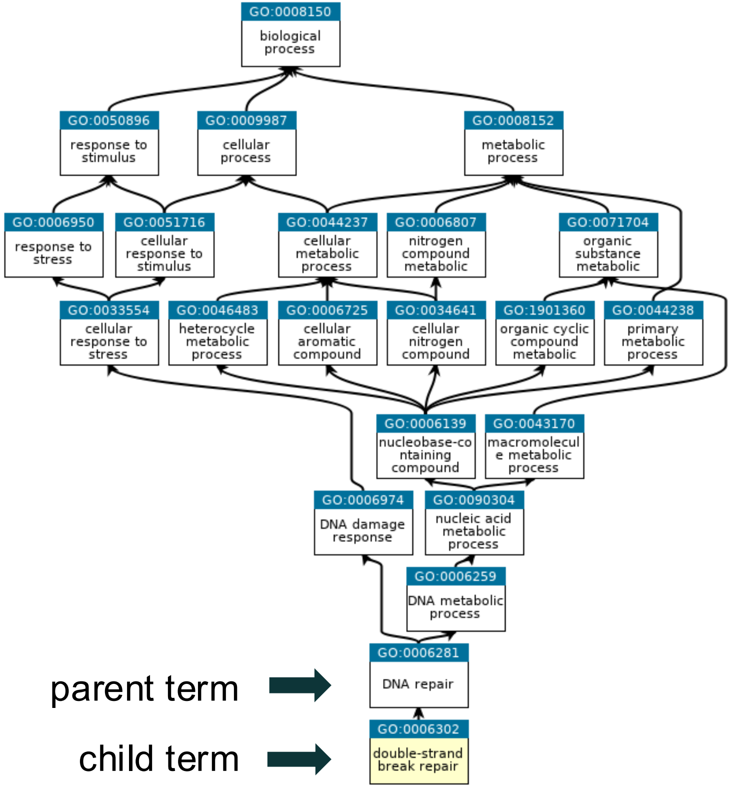
- Actually three separate ontologies:
  - **Molecular Function** – molecular level activities performed by gene products, e.g. *transporter activity* (broad) or *Toll-like receptor binding* (specific)
  - **Cellular Component** – the cellular location where a function is performed, e.g. *ribosome*
  - **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

root term

more genes

broad terms



GO:0008150
biological process
GO:0050896
response to stimulus
GO:0009987
cellular process
GO:0008152
metabolic process
GO:0006950
response to stress
GO:0051716
cellular response to stimulus
GO:0044237
cellular metabolic process
GO:0006807
nitrogen compound metabolic
GO:0071704
organic substance metabolic
GO:0033554
cellular response to stress
GO:0046483
heterocycle metabolic process
GO:0006725
cellular aromatic compound
GO:0034641
cellular nitrogen compound
GO:1901360
organic cyclic compound metabolic
GO:0044238
primary metabolic process
GO:0006139
nucleobase-containing compound
GO:0043170
macromolecule metabolic process
GO:0006974
DNA damage response
GO:0090304
nucleic acid metabolic process
GO:0006259
DNA metabolic process
GO:0006281
DNA repair
GO:0006302
double-strand break repair

parent term

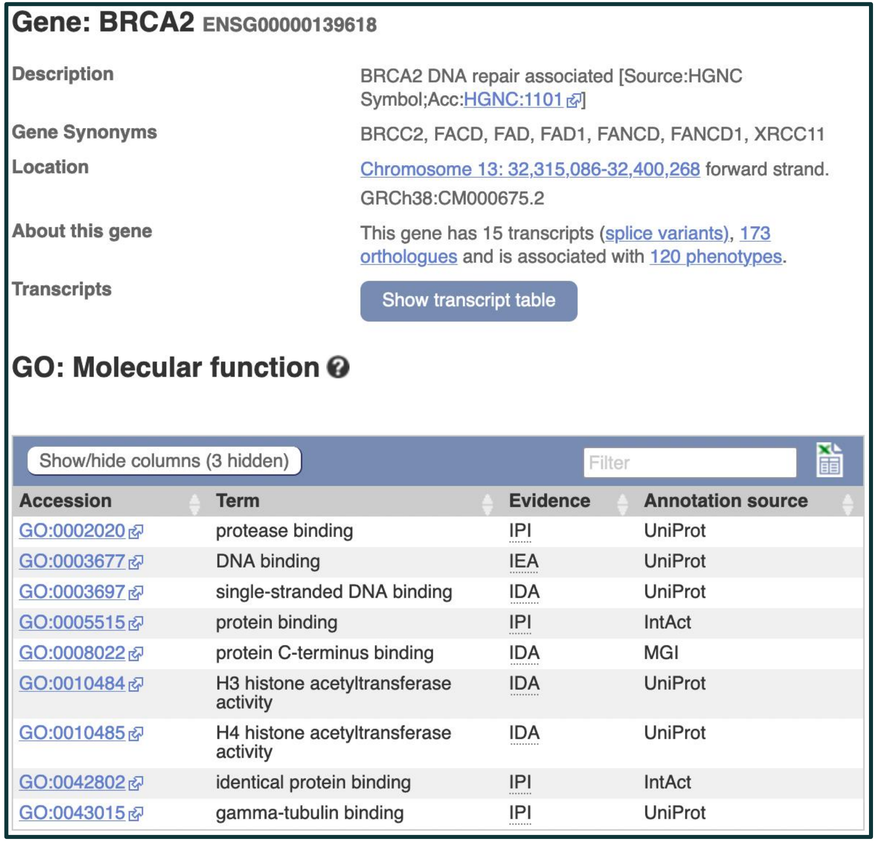
child term

QuickGO - https://www.ebi.ac.uk/QuickGO

fewer genes

specific terms

BRCA2 example



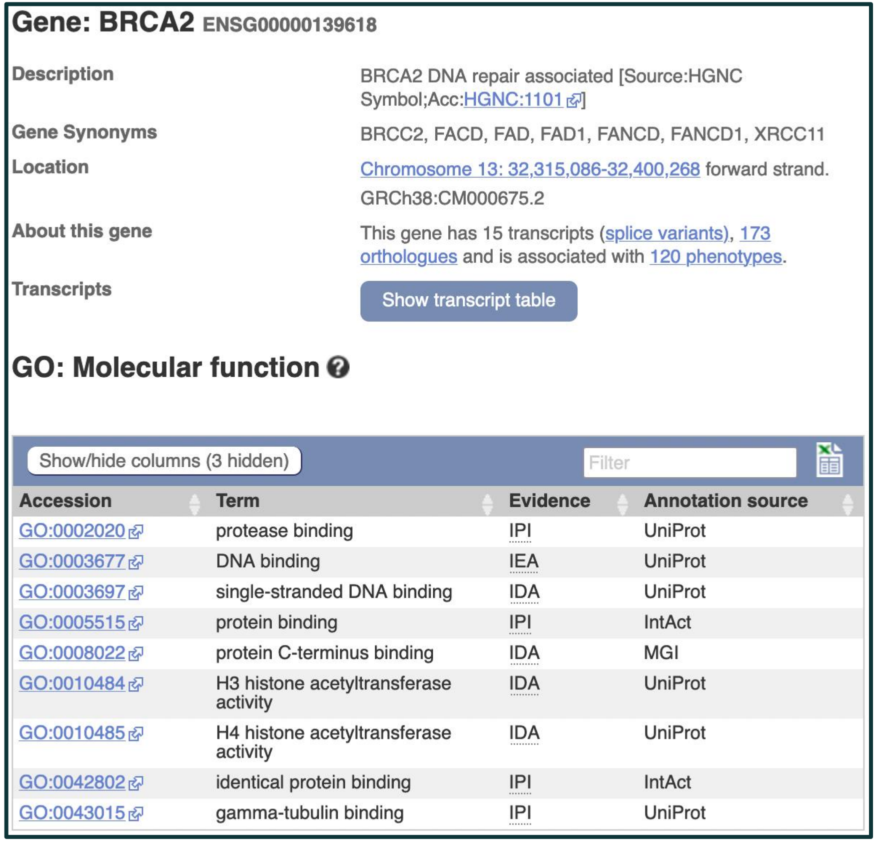
Gene: BRCA2 ENSG00000139618

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 (splice variants), 173 orthologues and is associated with 120 phenotypes.
Transcripts	Show transcript table

GO: Molecular function

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

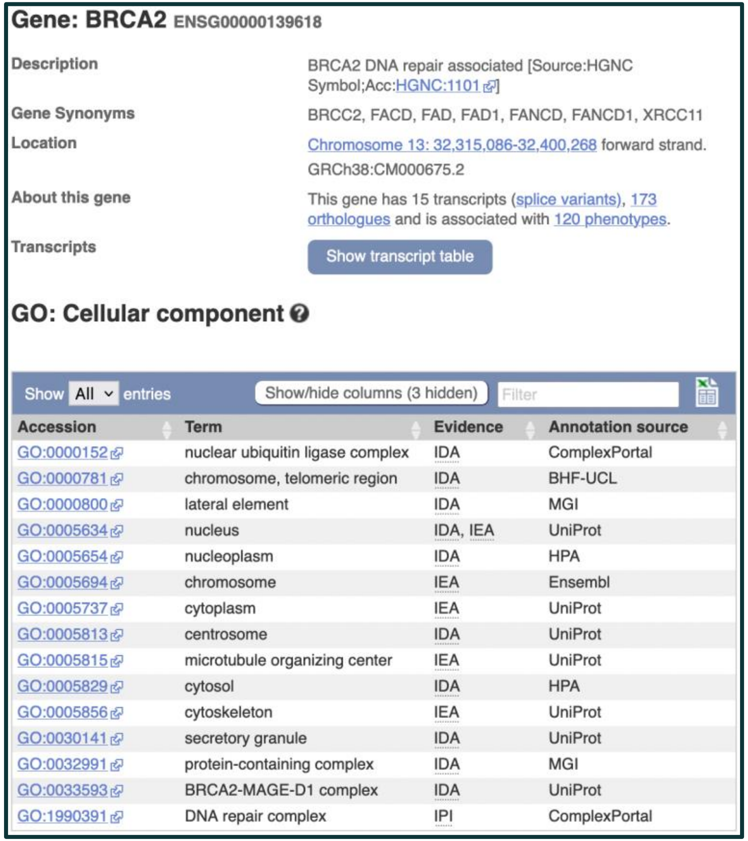


Gene: BRCA2 ENSG00000139618

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

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



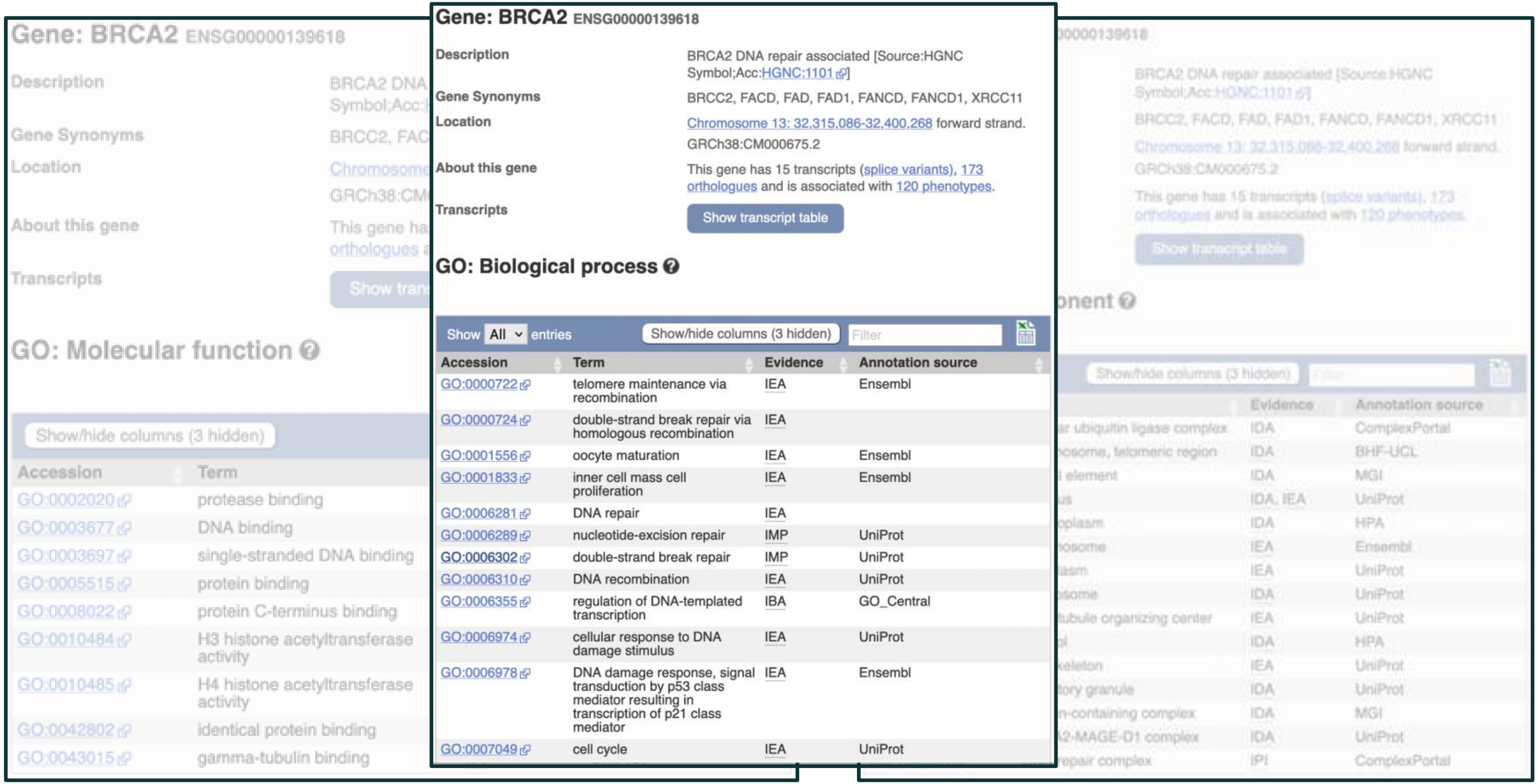
Gene: BRCA2 ENSG00000139618

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

Accession	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 ha
orthologues

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

onent

Show/hide columns (3 hidden) Filter

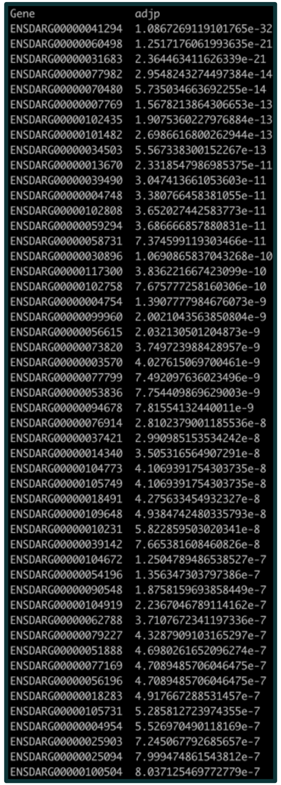
	Evidence	Annotation source
ar ubiquitin ligase complex	IDA	ComplexPortal
nosome, telomeric region	IDA	BHF-UCL
l element	IDA	MGI
us	IDA, IEA	UniProt
oplasm	IDA	HPA
nosome	IEA	Ensembl
lasm	IEA	UniProt
osome	IDA	UniProt
tubule organizing center	IEA	UniProt
ol	IDA	HPA
keleton	IEA	UniProt
tory granule	IDA	UniProt
n-containing complex	IDA	MGI
A2-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

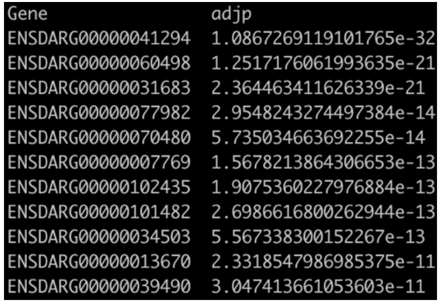
20,000 genes assayed



```
Gene                adjp
ENSDARG00000041294  1.0867269119101765e-32
ENSDARG00000060498  1.2517176061993635e-21
ENSDARG00000031683  2.364463411626339e-21
ENSDARG00000077982  2.9548243274497384e-14
ENSDARG00000070480  5.735034663692255e-14
ENSDARG00000007769  1.5678213864306653e-13
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
ENSDARG00000004748  3.380766458381055e-11
ENSDARG00000102808  3.652027442583773e-11
ENSDARG00000059294  3.686666857880831e-11
ENSDARG00000058731  7.374599119303466e-11
ENSDARG00000030896  1.0690865837043268e-10
ENSDARG00000117300  3.836221667423099e-10
ENSDARG00000102758  7.675777258160306e-10
ENSDARG00000004754  1.390777984676073e-9
ENSDARG00000099960  2.0021043563850804e-9
ENSDARG00000056615  2.032130501204873e-9
ENSDARG00000073820  3.749723988428957e-9
ENSDARG00000003570  4.027615069700461e-9
ENSDARG00000077799  7.492097636023496e-9
ENSDARG00000053836  7.754409869629003e-9
ENSDARG00000094678  7.81554132440011e-9
ENSDARG00000076914  2.8102379001185536e-8
ENSDARG00000037421  2.990985153534242e-8
ENSDARG00000014340  3.505316564907291e-8
ENSDARG00000104773  4.1069391754303735e-8
ENSDARG00000105749  4.1069391754303735e-8
ENSDARG00000018491  4.275633454932327e-8
ENSDARG00000109648  4.9384742480335793e-8
ENSDARG00000010231  5.822859503020341e-8
ENSDARG00000039142  7.665381608460826e-8
ENSDARG00000104672  1.2504789486538527e-7
ENSDARG00000054196  1.356347303797386e-7
ENSDARG00000090548  1.8758159693858449e-7
ENSDARG00000104919  2.2367046789114162e-7
ENSDARG00000062788  3.7107672341197336e-7
ENSDARG00000079227  4.3287909103165297e-7
ENSDARG00000051888  4.6980261652096274e-7
ENSDARG00000077169  4.7089485706046475e-7
ENSDARG00000056196  4.7089485706046475e-7
ENSDARG00000018283  4.917667288531457e-7
ENSDARG00000105731  5.285812723974355e-7
ENSDARG00000004954  5.526970490118169e-7
ENSDARG00000025903  7.245067792685657e-7
ENSDARG00000025094  7.999474861543812e-7
ENSDARG00000100504  8.037125469772779e-7
```

Adjusted
p-value
< 0.05

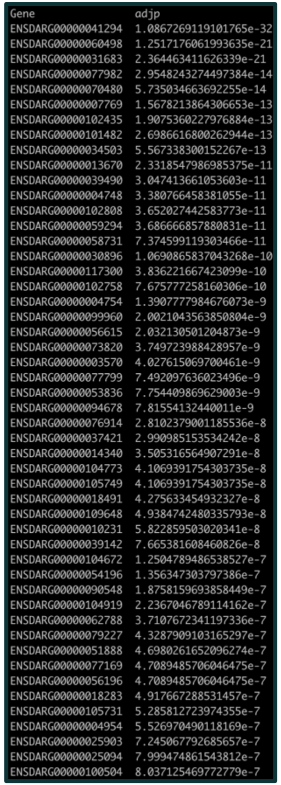
500 significantly DE genes



```
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
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
```

Testing for functional enrichment

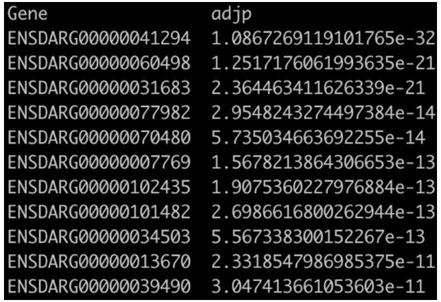
20,000 genes assayed



```
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
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
ENSDARG00000004748  3.380766458381055e-11
ENSDARG00000102808  3.652027442583773e-11
ENSDARG00000059294  3.686666857880831e-11
ENSDARG00000058731  7.374599119303466e-11
ENSDARG00000030896  1.0690865837043268e-10
ENSDARG00000117300  3.836221667423099e-10
ENSDARG00000102758  7.675777258160306e-10
ENSDARG00000004754  1.390777984676073e-9
ENSDARG00000099960  2.002104356385080 4e-9
ENSDARG00000056615  2.032130501204873e-9
ENSDARG00000073820  3.749723988428957e-9
ENSDARG00000003570  4.027615069700461e-9
ENSDARG00000077799  7.492097636023496e-9
ENSDARG00000053836  7.754409869629003e-9
ENSDARG00000094678  7.81554132440011e-9
ENSDARG00000076914  2.8102379001185536e-8
ENSDARG00000037421  2.990985153534242e-8
ENSDARG00000014340  3.505316564907291e-8
ENSDARG00000104773  4.1069391754303735e-8
ENSDARG00000105749  4.1069391754303735e-8
ENSDARG00000018491  4.275633454932327e-8
ENSDARG00000109648  4.9384742480335793e-8
ENSDARG00000010231  5.822859503020341e-8
ENSDARG00000039142  7.665381608460826e-8
ENSDARG00000104672  1.2504789486538527e-7
ENSDARG00000054196  1.356347303797386e-7
ENSDARG00000090548  1.8758159693858449e-7
ENSDARG00000104919  2.2367046789114162e-7
ENSDARG00000062788  3.7107672341197336e-7
ENSDARG00000079227  4.3287909103165297e-7
ENSDARG00000051888  4.6980261652096274e-7
ENSDARG00000077169  4.7089485706046475e-7
ENSDARG00000056196  4.7089485706046475e-7
ENSDARG00000018283  4.917667288531457e-7
ENSDARG00000105731  5.285812723974355e-7
ENSDARG00000004954  5.526970490118169e-7
ENSDARG00000025903  7.245067792685657e-7
ENSDARG00000025094  7.999474861543812e-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
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
```

200 genes annotated to DNA repair

200/500 = **40%**

(300 not annotated to DNA repair)

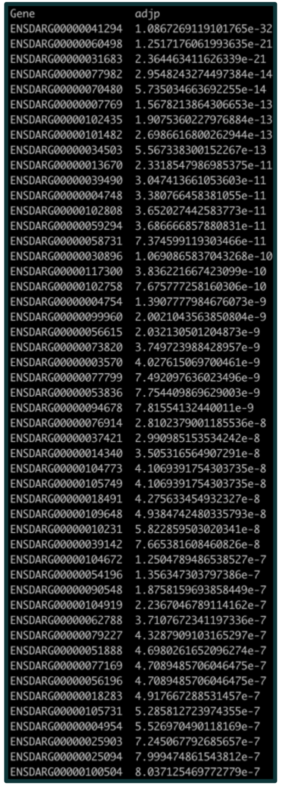
2000 genes annotated to function (e.g. DNA repair)

2000/20000 = **10%**

(18,000 not annotated to DNA repair)

Testing for functional enrichment

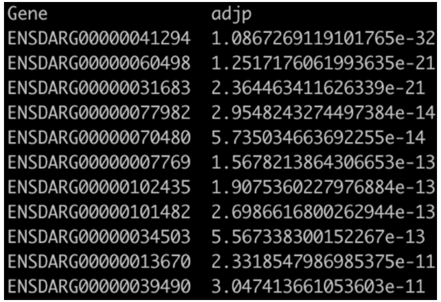
20,000 genes assayed



```
Gene                adjp
ENSDARG00000041294  1.0867269119101765e-32
ENSDARG00000060498  1.2517176061993635e-21
ENSDARG00000031683  2.364463411626339e-21
ENSDARG00000077982  2.9548243274497384e-14
ENSDARG00000070480  5.735034663692255e-14
ENSDARG00000007769  1.5678213864306653e-13
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
ENSDARG00000004748  3.380766458381055e-11
ENSDARG00000102808  3.652027442583773e-11
ENSDARG00000059294  3.686666857880831e-11
ENSDARG00000058731  7.374599119303466e-11
ENSDARG00000030896  1.0690865837043268e-10
ENSDARG00000117300  3.836221667423099e-10
ENSDARG00000102758  7.675777258160306e-10
ENSDARG00000004754  1.390777984676073e-9
ENSDARG00000099960  2.002104356385080 4e-9
ENSDARG00000056615  2.032130501204873e-9
ENSDARG00000073820  3.749723988428957e-9
ENSDARG00000003570  4.027615069700461e-9
ENSDARG00000077799  7.492097636023496e-9
ENSDARG00000053836  7.754409869629003e-9
ENSDARG00000094678  7.81554132440011e-9
ENSDARG00000076914  2.8102379001185536e-8
ENSDARG00000037421  2.990985153534242e-8
ENSDARG00000014340  3.505316564907291e-8
ENSDARG00000104773  4.1069391754303735e-8
ENSDARG00000105749  4.1069391754303735e-8
ENSDARG00000018491  4.275633454932327e-8
ENSDARG00000109648  4.9384742480335793e-8
ENSDARG00000010231  5.822859503020341e-8
ENSDARG00000039142  7.665381608460826e-8
ENSDARG00000104672  1.2504789486538527e-7
ENSDARG00000054196  1.356347303797386e-7
ENSDARG00000090548  1.8758159693858449e-7
ENSDARG00000104919  2.2367046789114162e-7
ENSDARG00000062788  3.7107672341197336e-7
ENSDARG00000079227  4.3287909103165297e-7
ENSDARG00000051888  4.6980261652096274e-7
ENSDARG00000077169  4.7089485706046475e-7
ENSDARG00000056196  4.7089485706046475e-7
ENSDARG00000018283  4.917667288531457e-7
ENSDARG00000105731  5.285812723974355e-7
ENSDARG00000004954  5.526970490118169e-7
ENSDARG00000025903  7.245067792685657e-7
ENSDARG00000025094  7.999474861543812e-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
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-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 to function (e.g. DNA repair)

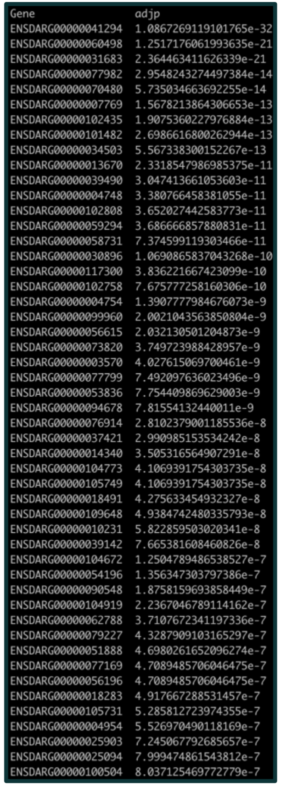
2000/20000 = **10%**

(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

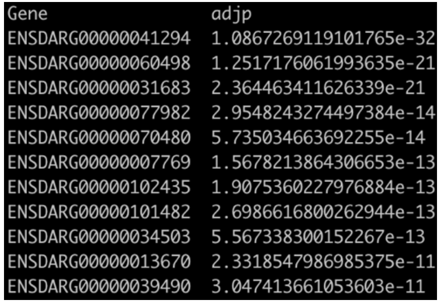
20,000 genes assayed



```
Gene               adjp
ENSDARG00000041294  1.0867269119101765e-32
ENSDARG00000060498  1.2517176061993635e-21
ENSDARG00000031683  2.364463411626339e-21
ENSDARG00000077982  2.9548243274497384e-14
ENSDARG00000070480  5.735034663692255e-14
ENSDARG00000007769  1.5678213864306653e-13
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-13
ENSDARG00000013670  2.3318547986985375e-11
ENSDARG00000039490  3.047413661053603e-11
ENSDARG00000004748  3.380766458381055e-11
ENSDARG00000102808  3.652027442583773e-11
ENSDARG00000059294  3.686666857880831e-11
ENSDARG00000058731  7.374599119303466e-11
ENSDARG00000030896  1.0690865837043268e-10
ENSDARG00000117300  3.836221667423099e-10
ENSDARG00000102758  7.675777258160306e-10
ENSDARG00000004754  1.3907777984676073e-9
ENSDARG00000099960  2.002104356385080 4e-9
ENSDARG00000056615  2.032130501204873e-9
ENSDARG00000073820  3.749723988428957e-9
ENSDARG00000003570  4.027615069700461e-9
ENSDARG00000077799  7.492097636023496e-9
ENSDARG00000053836  7.754409869629003e-9
ENSDARG00000094678  7.81554132440011e-9
ENSDARG00000076914  2.8102379001185536e-8
ENSDARG00000037421  2.990985153534242e-8
ENSDARG00000014340  3.505316564907291e-8
ENSDARG00000104773  4.1069391754303735e-8
ENSDARG00000105749  4.1069391754303735e-8
ENSDARG00000018491  4.275633454932327e-8
ENSDARG00000109648  4.9384742480335793e-8
ENSDARG00000010231  5.822859503020341e-8
ENSDARG00000039142  7.665381608460826e-8
ENSDARG00000104672  1.2504789486538527e-7
ENSDARG00000054196  1.356347303797386e-7
ENSDARG00000090548  1.8758159693858449e-7
ENSDARG00000104919  2.2367046789114162e-7
ENSDARG00000062788  3.7107672341197336e-7
ENSDARG00000079227  4.3287909103165297e-7
ENSDARG00000051888  4.6980261652096274e-7
ENSDARG00000077169  4.7089485706046475e-7
ENSDARG00000056196  4.7089485706046475e-7
ENSDARG00000018283  4.917667288531457e-7
ENSDARG00000105731  5.285812723974355e-7
ENSDARG00000004954  5.526970490118169e-7
ENSDARG00000025903  7.245067792685657e-7
ENSDARG00000025094  7.999474861543812e-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
ENSDARG00000102435  1.9075360227976884e-13
ENSDARG00000101482  2.6986616800262944e-13
ENSDARG00000034503  5.567338300152267e-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 annota
function (e.g. DNA

2000/20000 = **10%**

(18,000 not annota
DNA repair)

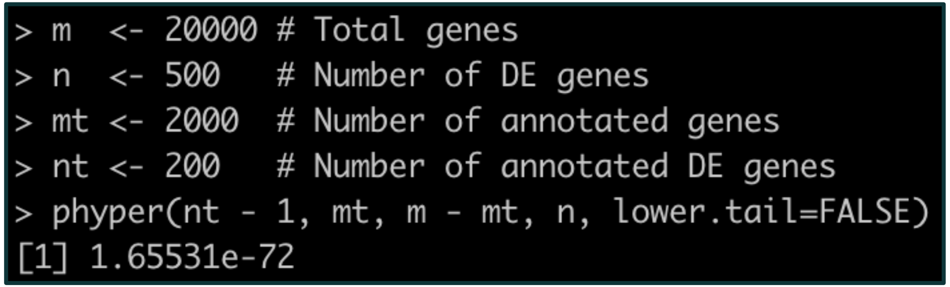
	DE	Not DE	**Total**
Annotated to DNA repair	200	1800	2000
Not annotated to DNA repair	300	17700	18000
Total	500	19500	20000

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

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

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



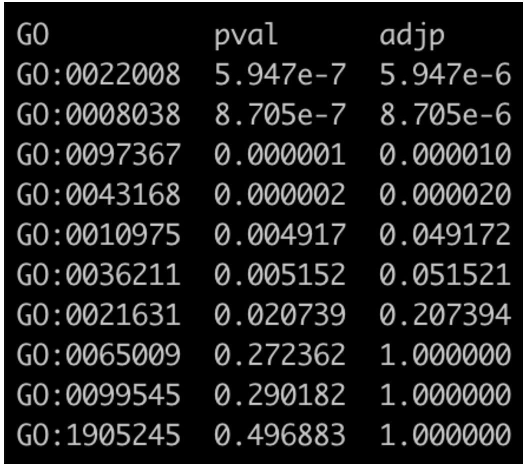
```
> m  <- 20000 # Total genes
> n  <- 500   # Number of DE genes
> mt <- 2000  # Number of annotated genes
> nt <- 200   # Number of annotated DE genes
> phyper(nt - 1, mt, m - mt, n, lower.tail=FALSE)
[1] 1.65531e-72
```

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



```
GO          pval      adjp
GO:0022008  5.947e-7  5.947e-6
GO:0008038  8.705e-7  8.705e-6
GO:0097367  0.000001  0.000010
GO:0043168  0.000002  0.000020
GO:0010975  0.004917  0.049172
GO:0036211  0.005152  0.051521
GO:0021631  0.020739  0.207394
GO:0065009  0.272362  1.000000
GO:0099545  0.290182  1.000000
GO:1905245  0.496883  1.000000
```

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!

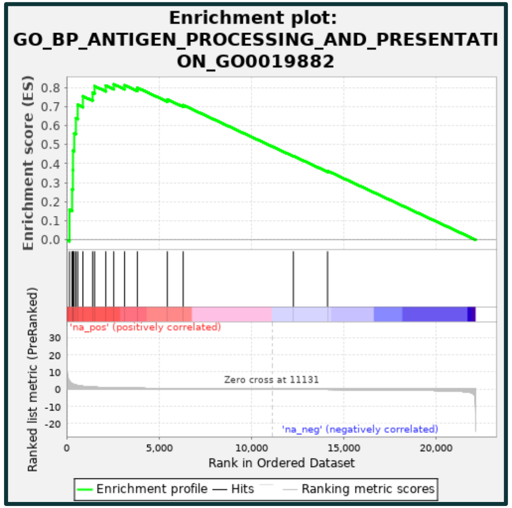
- Quiz on Mentimeter (www.menti.com)

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 800 species/strains/varieties

Other functional enrichment tools

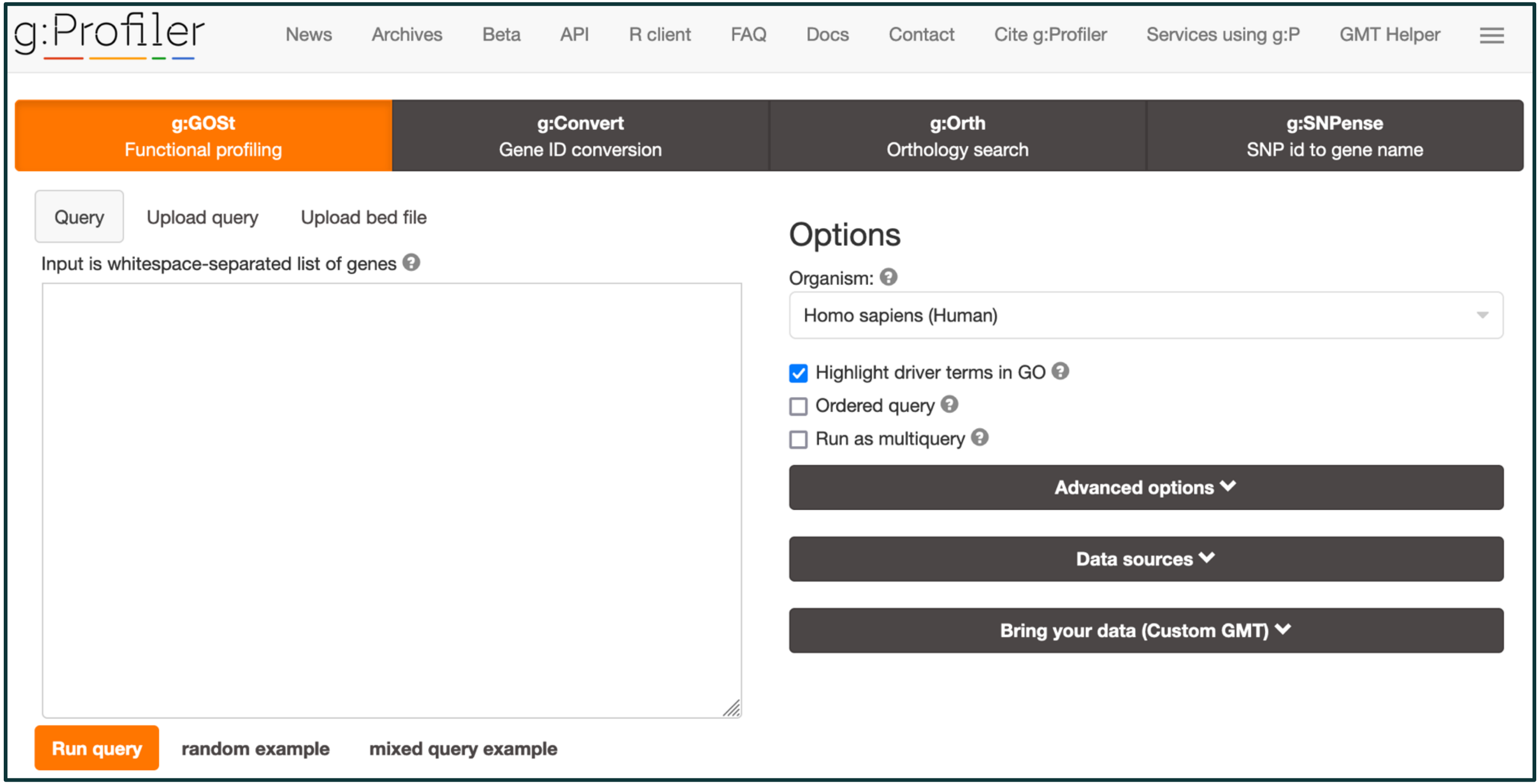
- Other tools are available (and good):
  - Enrichr (maayanlab.cloud/Enrichr/):
    - Web-based
    - Similar to g:Profiler
    - Only human, mouse, fly, yeast, worm and zebrafish
  - GSEA (www.gsea-msigdb.org/gsea/):
    - Desktop software
    - Implements GSEA method
    - Works on whole genome ranked 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 July last year, g:Profiler had been using Ensembl 113, which came out in October 2024
- Recommend using Ensembl IDs as input, but not essential

g:Profiler



Input is whitespace-separated list of genes

Options

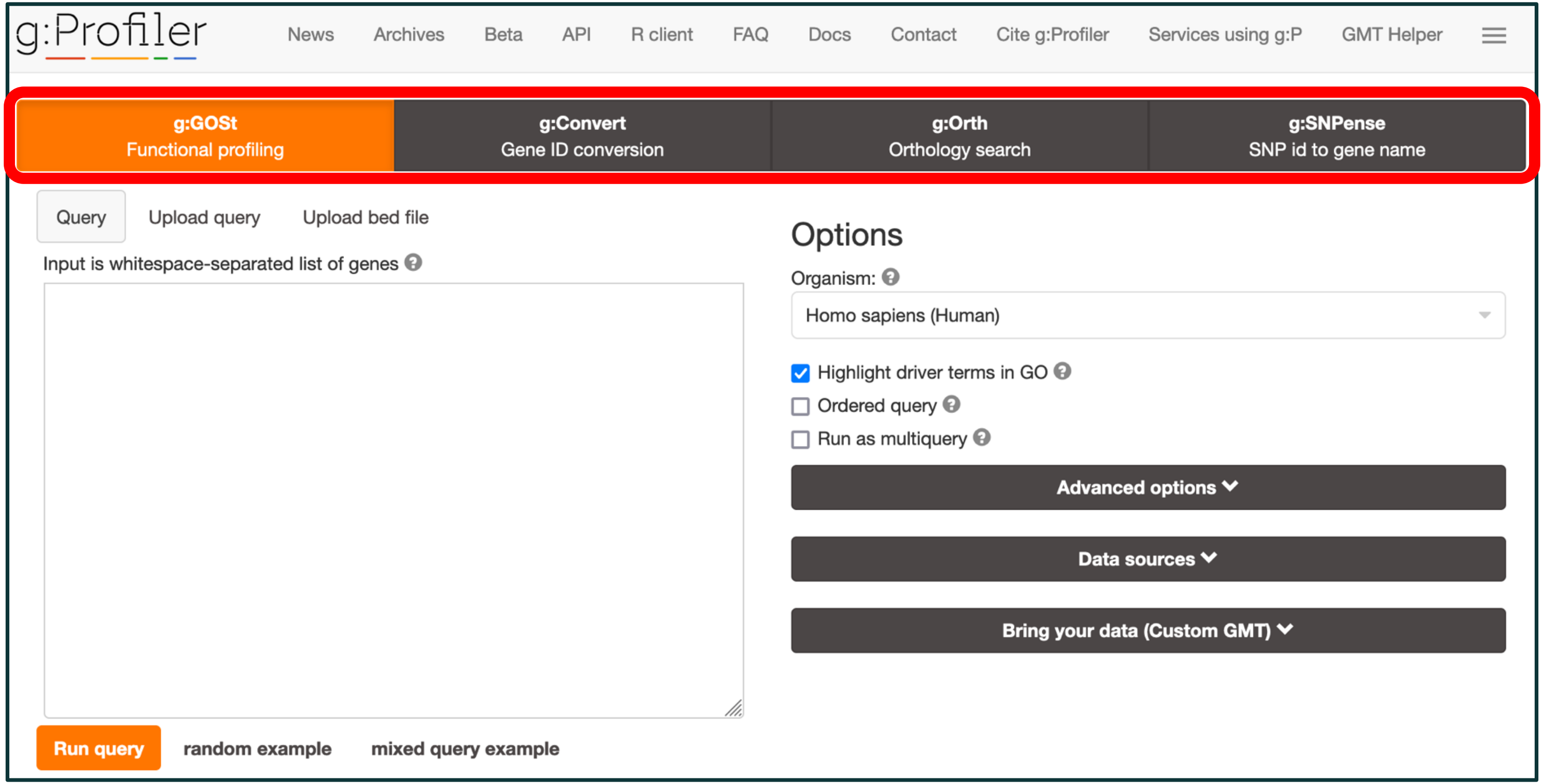
Organism: Homo sapiens (Human)

☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

g:Profiler – four tools



Input is whitespace-separated list of genes

Options

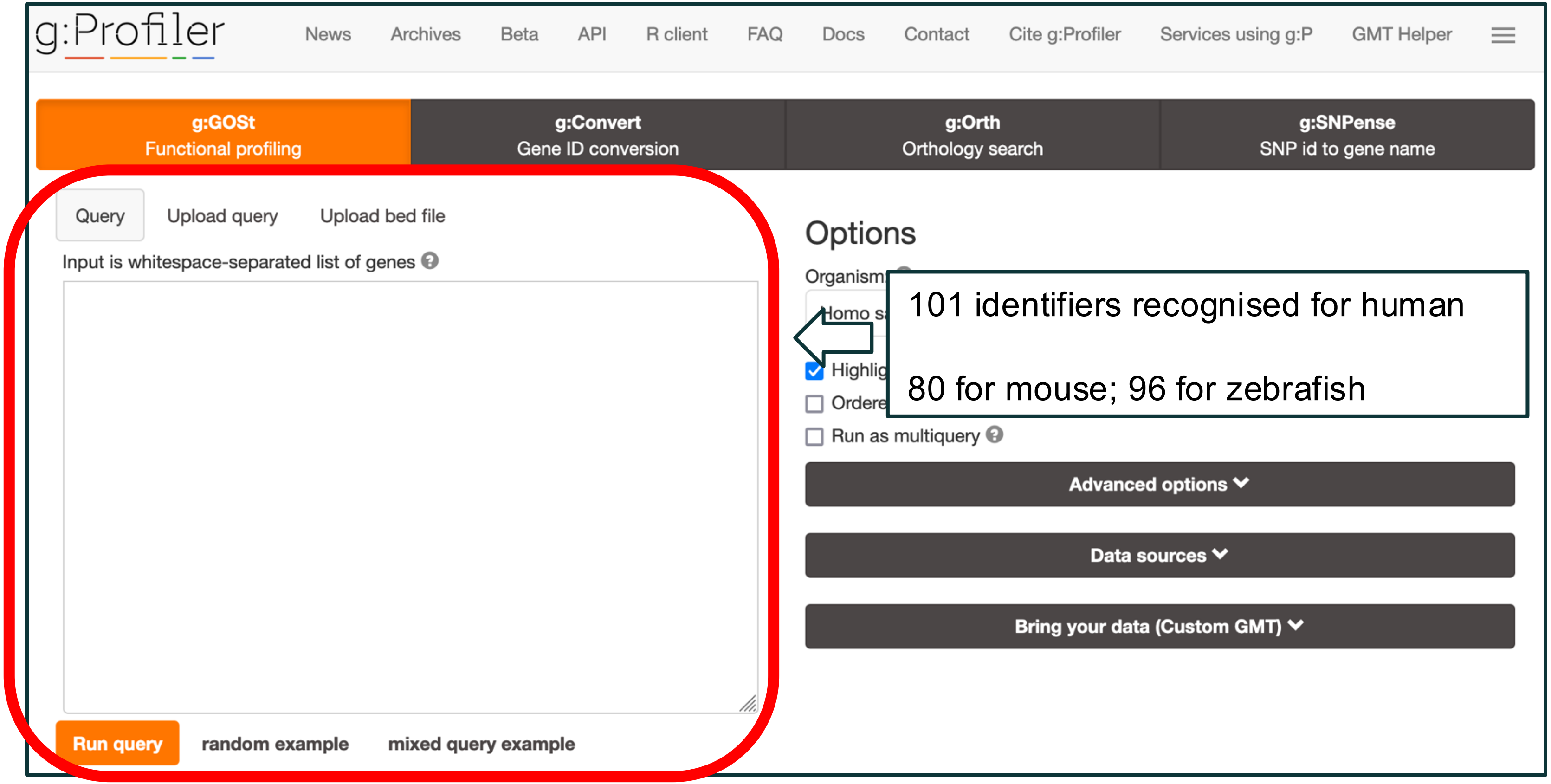
Organism: Homo sapiens (Human)

☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

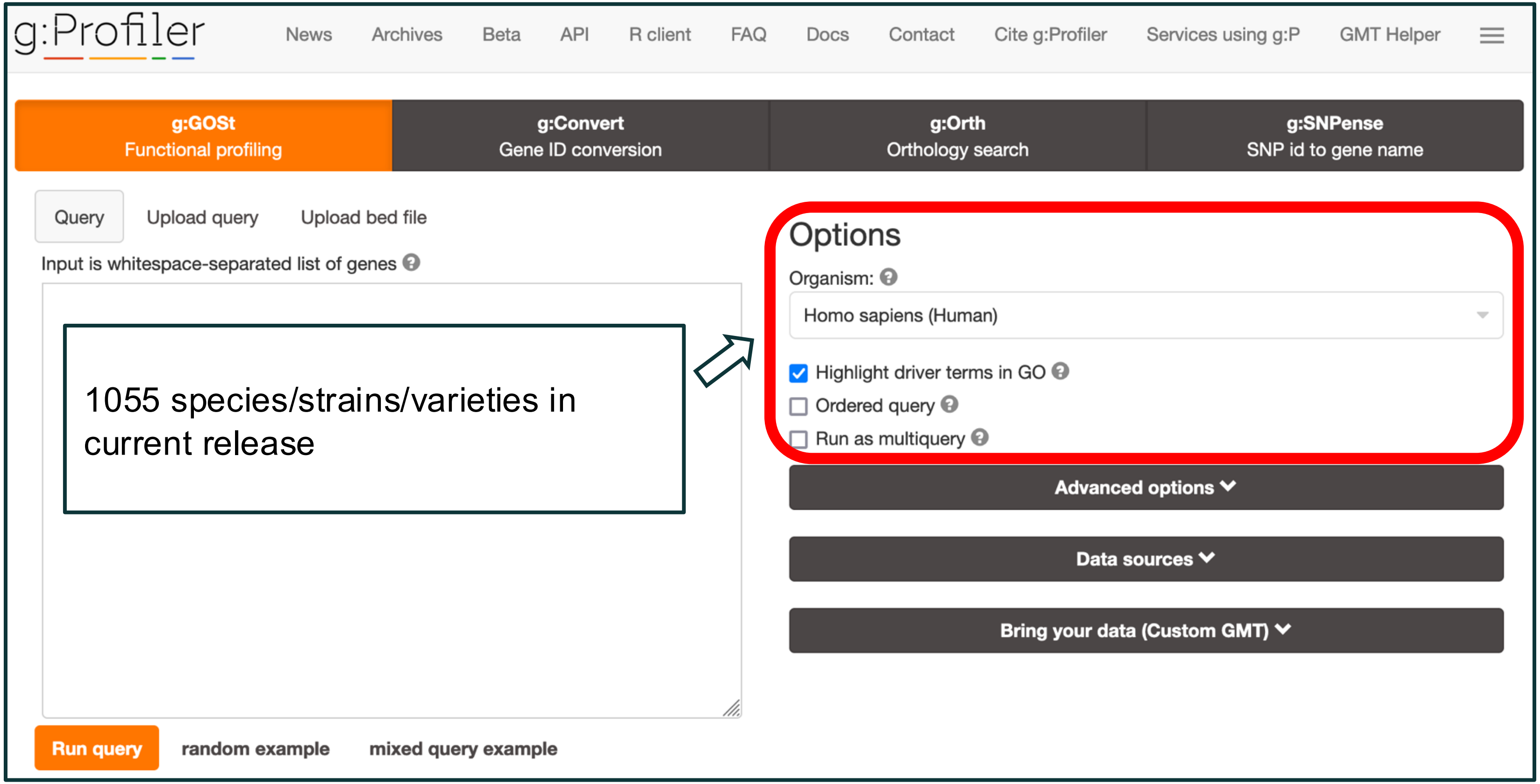
g:Profiler – gene list



101 identifiers recognised for human

80 for mouse; 96 for zebrafish

g:Profiler – options



Input is whitespace-separated list of genes

1055 species/strains/varieties in current release

Options

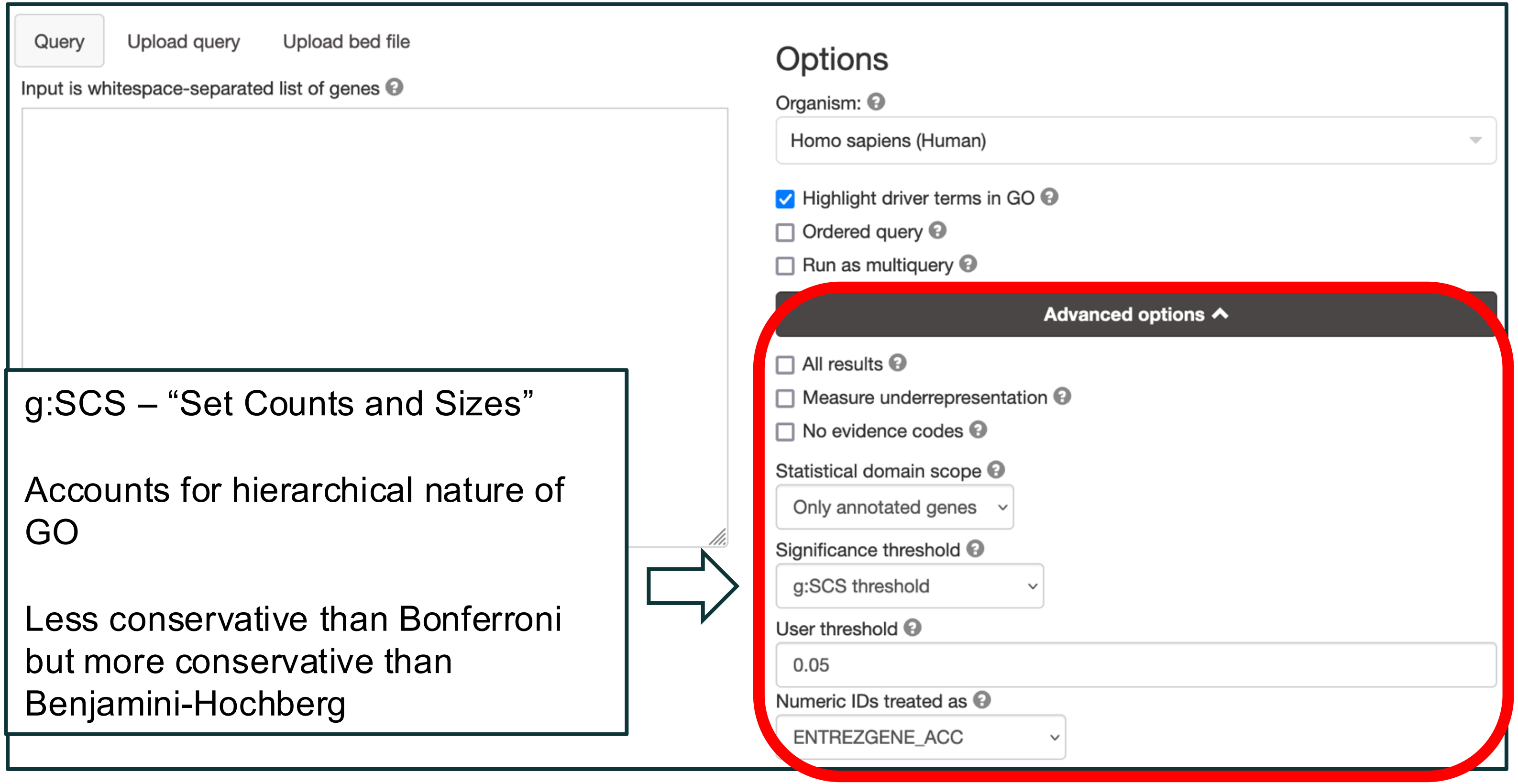
Organism: Homo sapiens (Human)

☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

g:Profiler – advanced options



Input is whitespace-separated list of genes

Options

Organism: Homo sapiens (Human)

☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

□ 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:SCS – “Set Counts and Sizes”

Accounts for hierarchical nature of GO

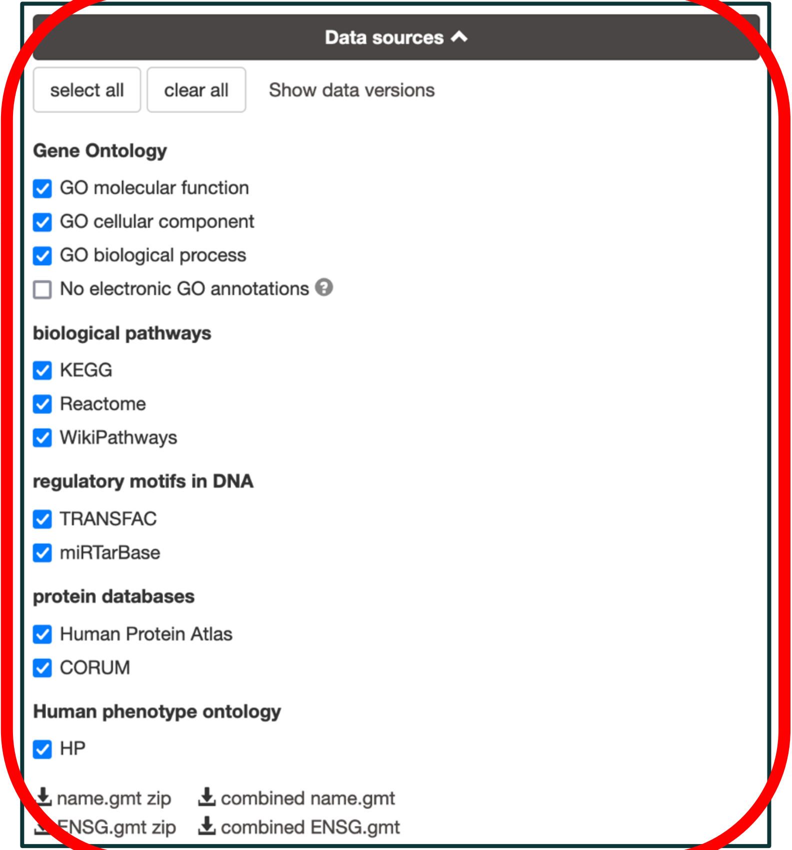
Less conservative than Bonferroni but more conservative than Benjamini-Hochberg

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

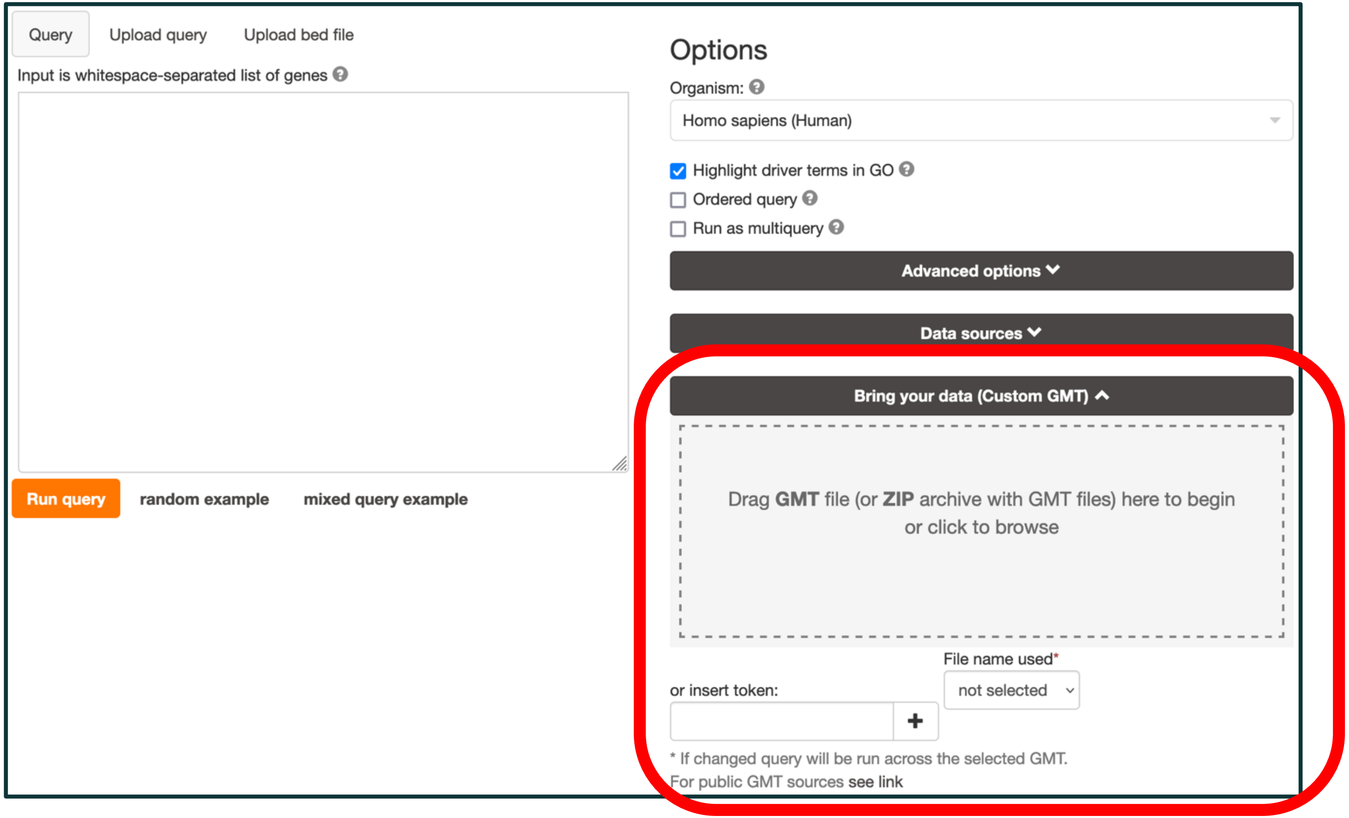
name.gmt zip | combined name.gmt
ENSG.gmt zip | combined ENSG.gmt

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



Input is whitespace-separated list of genes

Options

Organism: Homo sapiens (Human)

☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

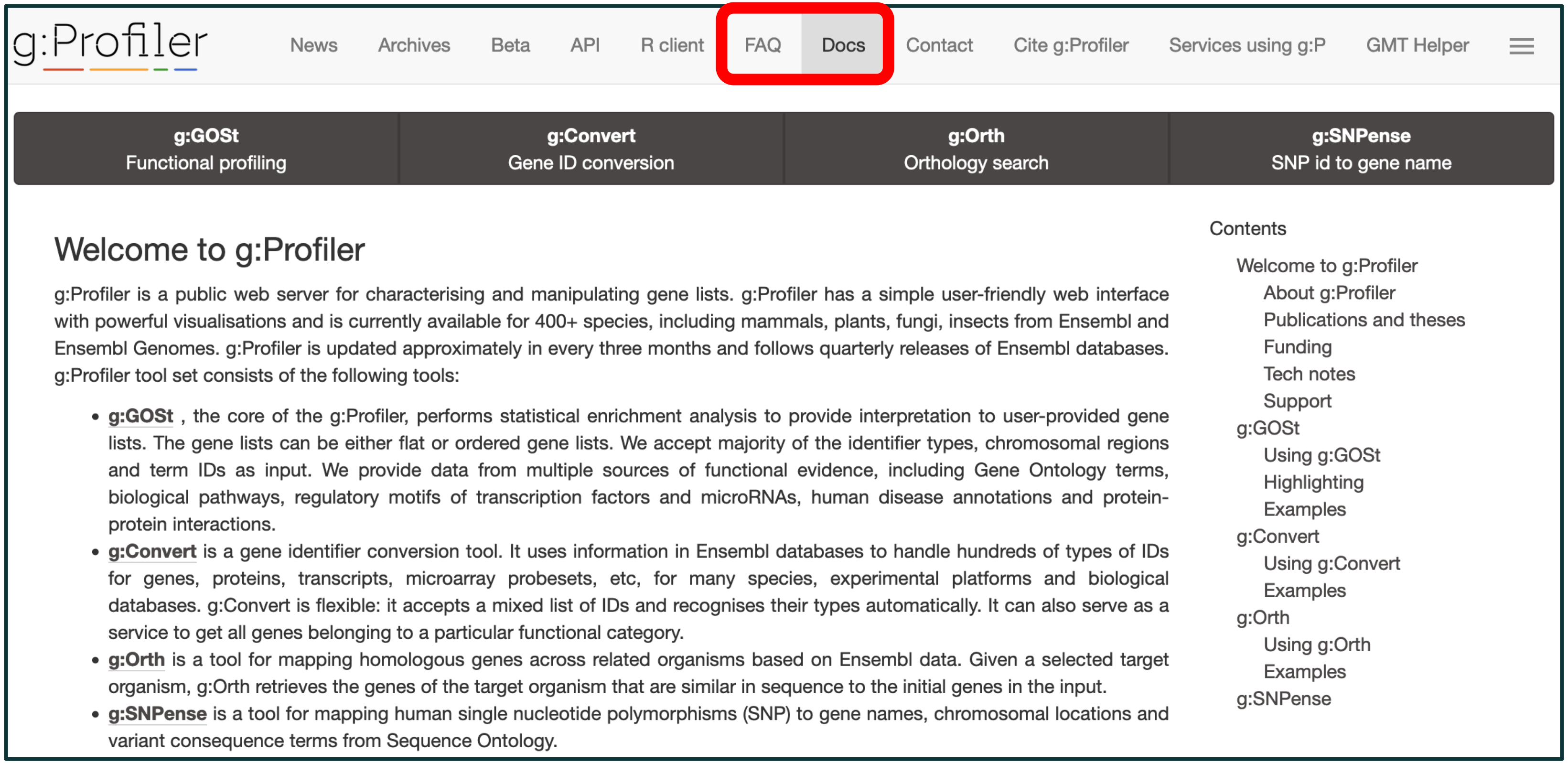
Drag **GMT** file (or **ZIP** archive with GMT files) here to begin or click to browse

or insert token:

File name used*: not selected

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

g:Profiler – documentation



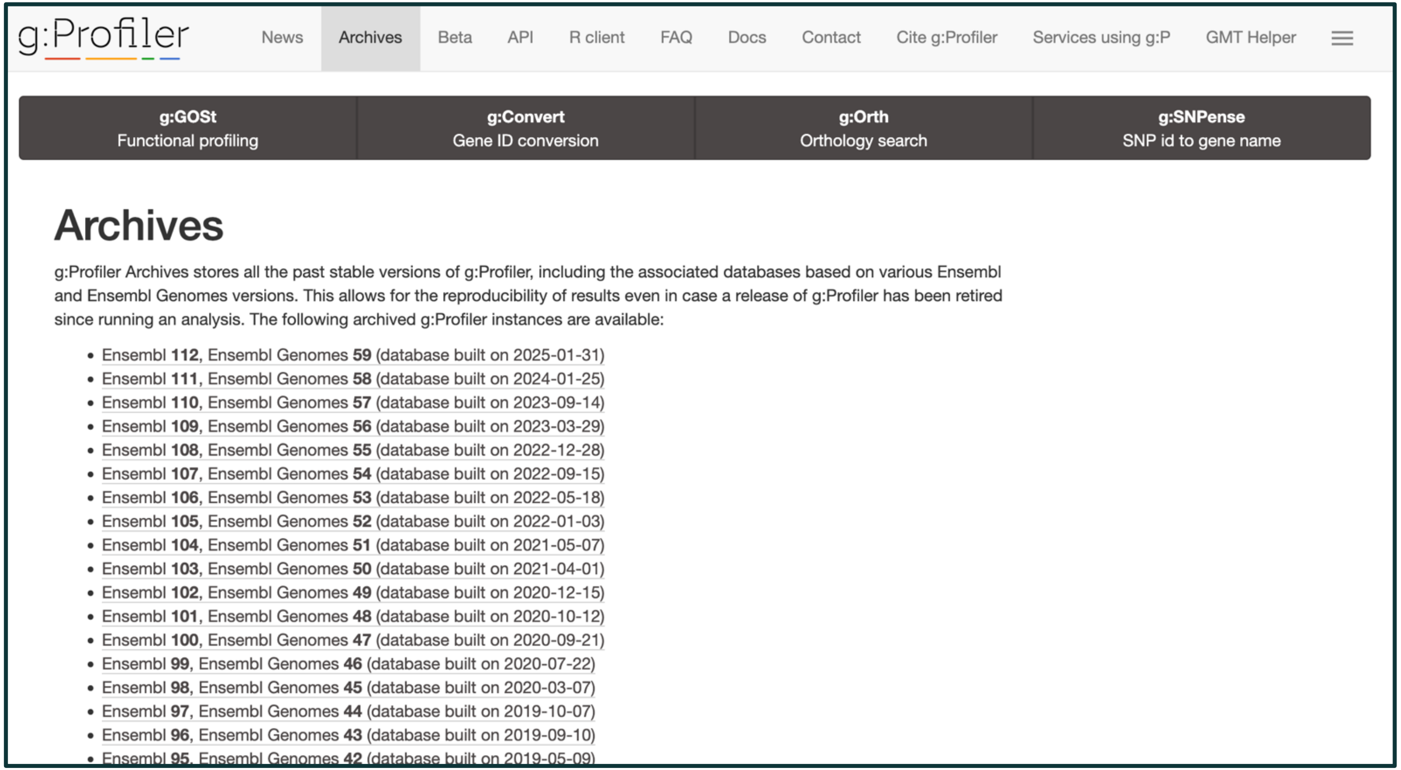
Welcome to g:Profiler

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

g:Profiler – archives

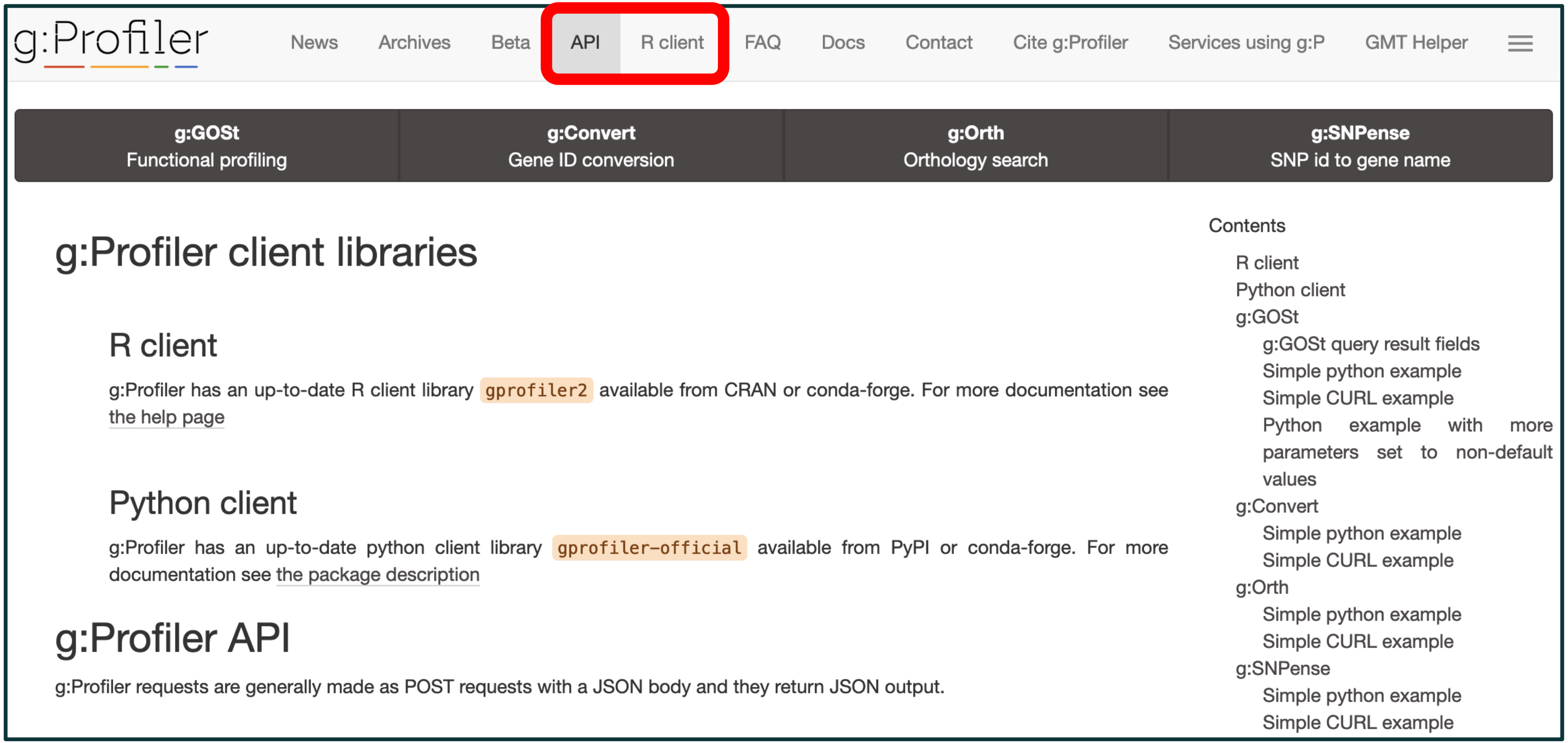


Archives

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 **112**, Ensembl Genomes **59** (database built on 2025-01-31)
- Ensembl **111**, Ensembl Genomes **58** (database built on 2024-01-25)
- 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)

g:Profiler – API and libraries



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

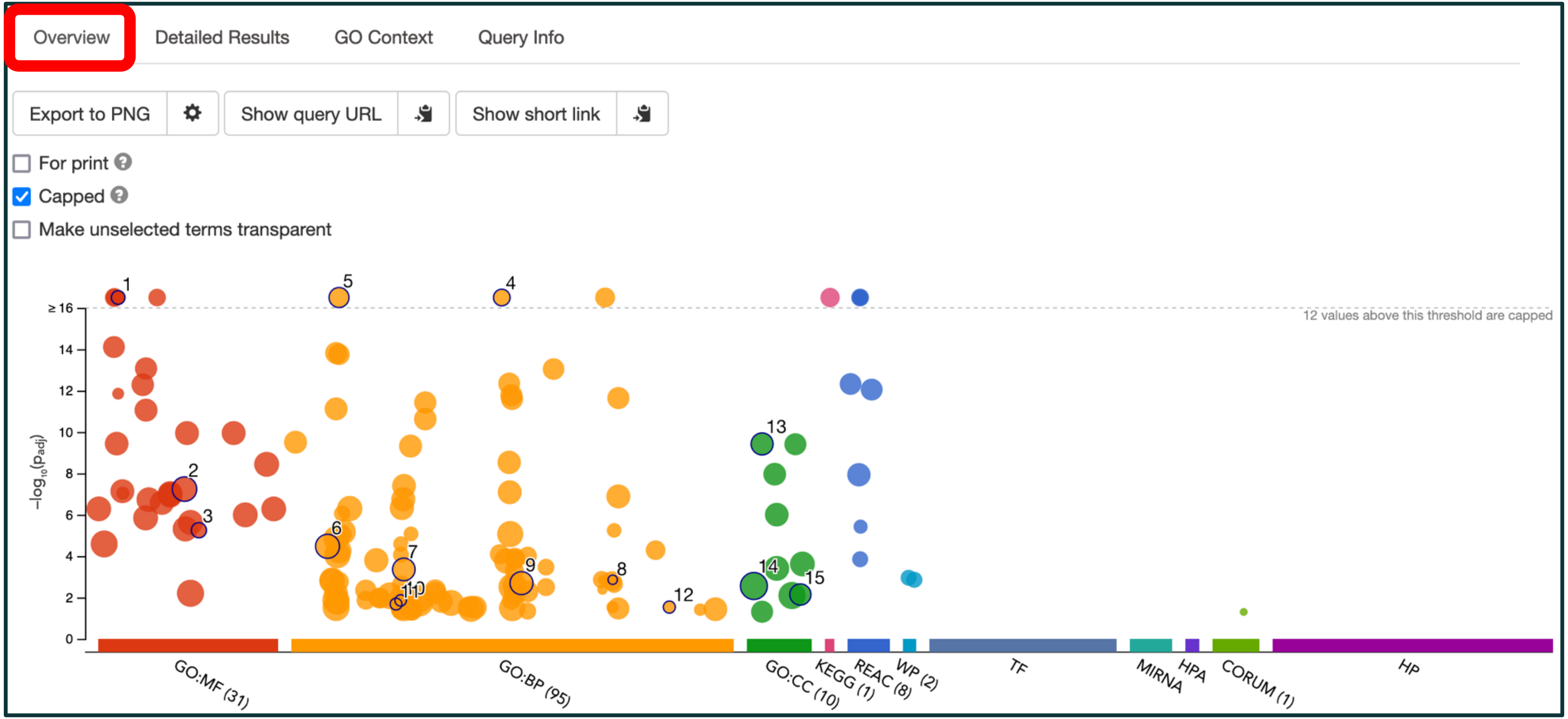
g:Profiler has an up-to-date python client library `gprofiler-official` available from PyPI or conda-forge. For more documentation see the package description

g:Profiler API

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

Contents

g:Profiler – overview

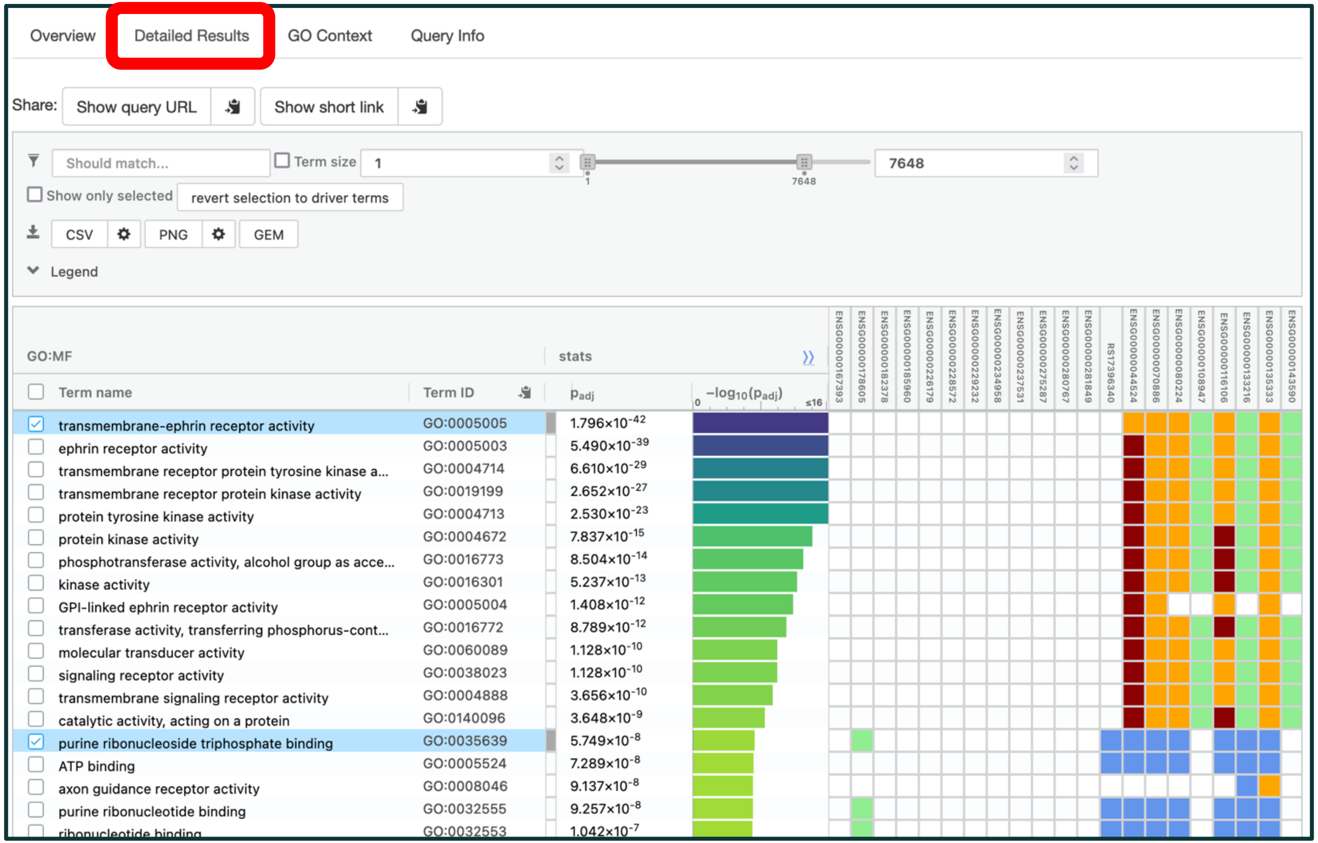


☐ For print

☑ Capped

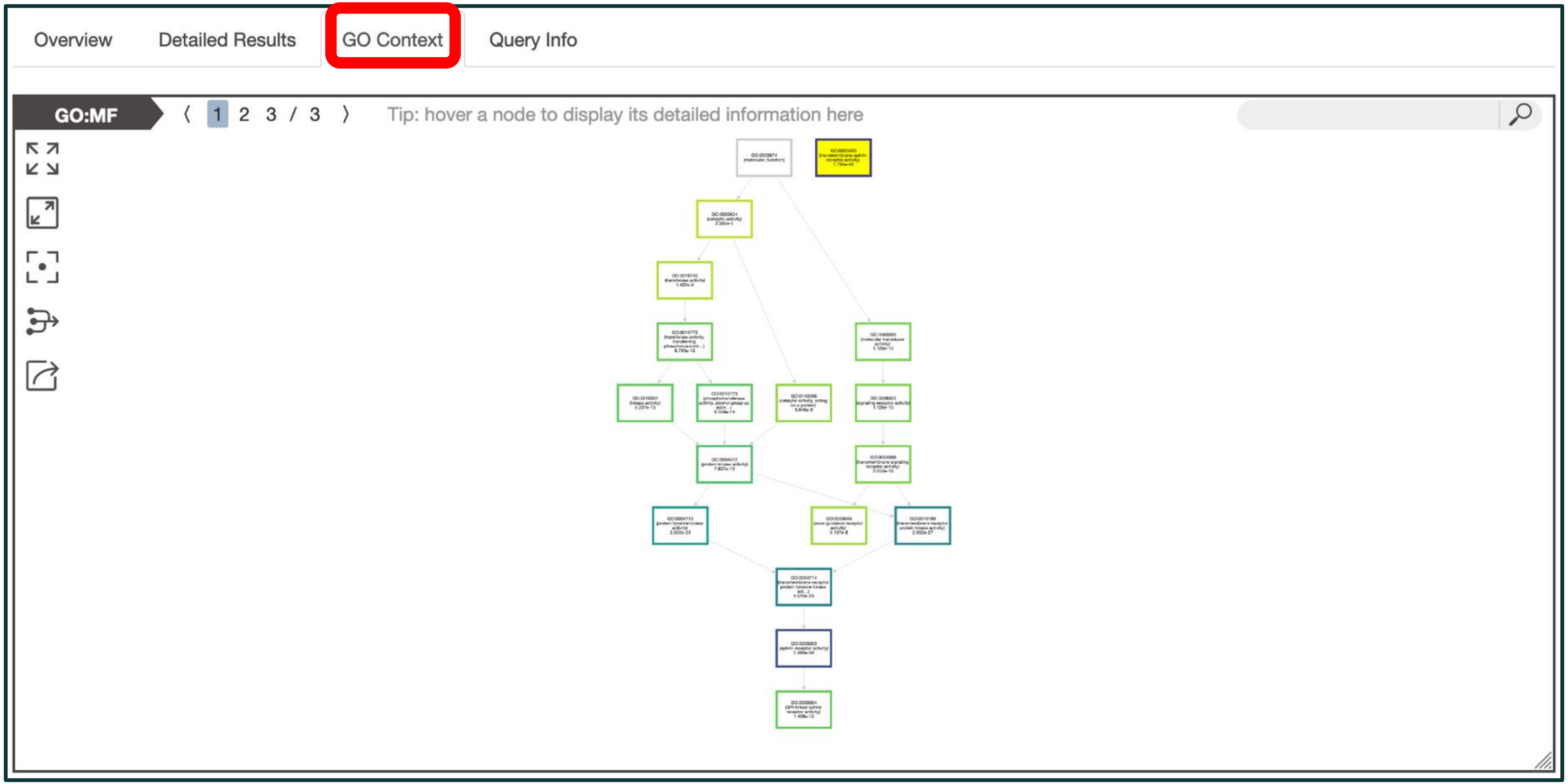
☐ Make unselected terms transparent

g:Profiler – detailed results

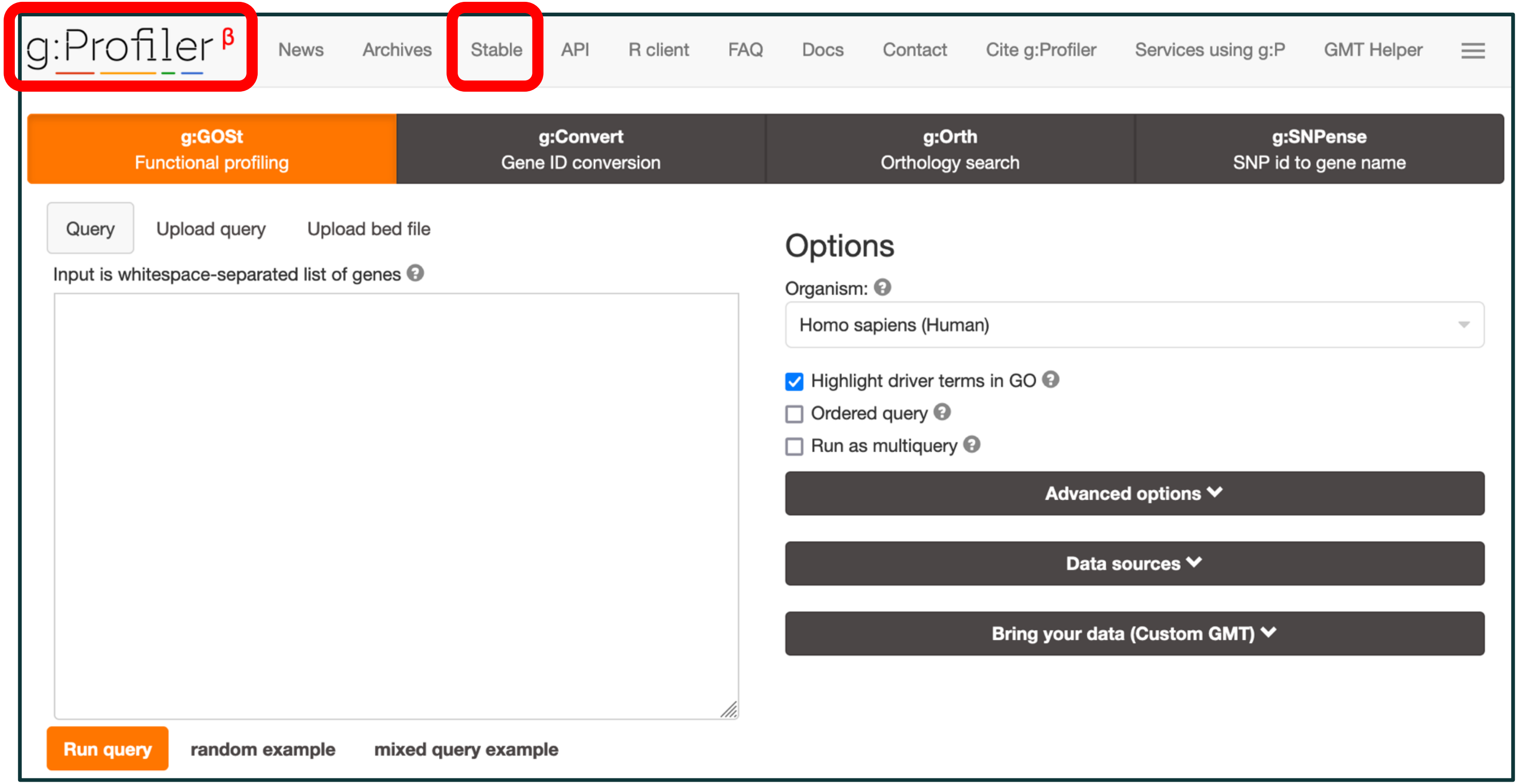


□	Term name	Term ID	p_{adj}
☑	transmembrane-ephrin receptor activity	GO:0005005	1.796×10^{-42}
□	ephrin receptor activity	GO:0005003	5.490×10^{-39}
□	transmembrane receptor protein tyrosine kinase a...	GO:0004714	6.610×10^{-29}
□	transmembrane receptor protein kinase activity	GO:0019199	2.652×10^{-27}
□	protein tyrosine kinase activity	GO:0004713	2.530×10^{-23}
□	protein kinase activity	GO:0004672	7.837×10^{-15}
□	phosphotransferase activity, alcohol group as acce...	GO:0016773	8.504×10^{-14}
□	kinase activity	GO:0016301	5.237×10^{-13}
□	GPI-linked ephrin receptor activity	GO:0005004	1.408×10^{-12}
□	transferase activity, transferring phosphorus-cont...	GO:0016772	8.789×10^{-12}
□	molecular transducer activity	GO:0060089	1.128×10^{-10}
□	signaling receptor activity	GO:0038023	1.128×10^{-10}
□	transmembrane signaling receptor activity	GO:0004888	3.656×10^{-10}
□	catalytic activity, acting on a protein	GO:0140096	3.648×10^{-9}
☑	purine ribonucleoside triphosphate binding	GO:0035639	5.749×10^{-8}
□	ATP binding	GO:0005524	7.289×10^{-8}
□	axon guidance receptor activity	GO:0008046	9.137×10^{-8}
□	purine ribonucleotide binding	GO:0032555	9.257×10^{-8}
□	ribonucleotide binding	GO:0032553	1.042×10^{-7}

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
GO:0003674 (molecular_function)
GO:0005005 (transmembrane-ephrin receptor activity) 1.795e-42
GO:0003824 (catalytic activity) 2.582e-5
GO:0016740 (transferase activity) 1.425e-6
GO:0016772 (transferase activity, transferring phosphorus-cont...) 8.795e-12
GO:0060089 (molecular transducer activity) 1.128e-10
GO:0016301 (kinase activity) 5.237e-13
GO:0016773 (phosphotransferase activity, alcohol group as acce...) 9.504e-14
GO:0140096 (catalytic activity, acting on a protein) 3.618e-9
GO:0038023 (signaling receptor activity) 1.128e-10
GO:0004672 (protein kinase activity) 7.837e-15
GO:0004888 (transmembrane signaling receptor activity) 3.656e-10
GO:0004713 (protein tyrosine kinase activity) 2.530e-23
GO:0008046 (axon guidance receptor activity) 5.127e-8
GO:0019199 (transmembrane receptor protein kinase activity) 2.652e-27
GO:0004714 (transmembrane receptor protein tyrosine kinase act...) 6.610e-29
GO:0005003 (ephrin receptor activity) 5.490e-39
GO:0005004 (GPI-linked ephrin receptor activity) 1.408e-12

g:Profiler – beta



g:Profiler β

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

Input is whitespace-separated list of genes

Options

Organism: Homo sapiens (Human)

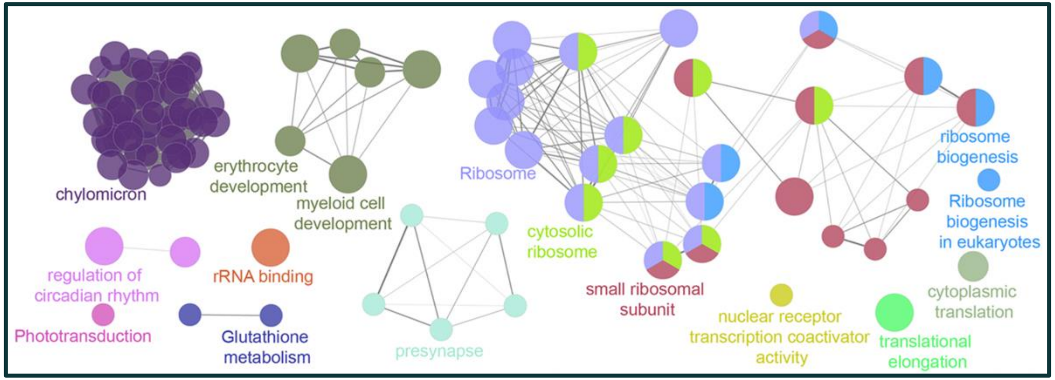
☑ Highlight driver terms in GO

□ Ordered query

□ Run as multiquery

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:

  rnaseq2026.buschlab.org

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