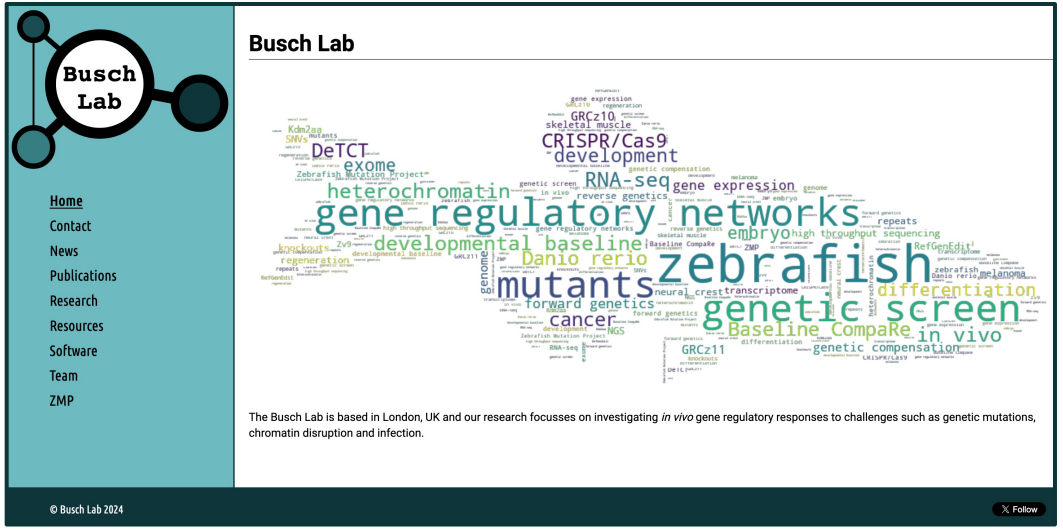
Introduction to RNA-seq and functional interpretation:

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

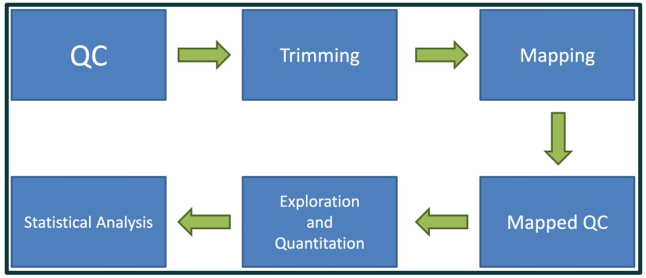


Questions

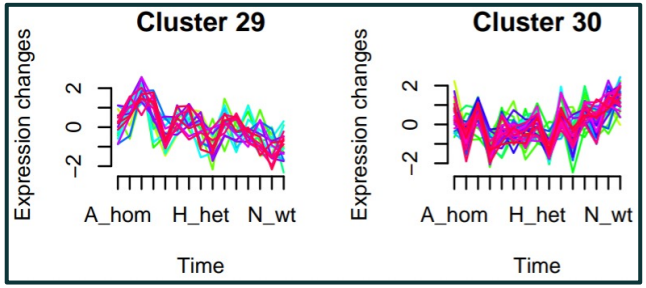
- For urgent questions, either:
  - Use Zoom’s Chat
  - Unmute and ask
- If you can wait, then ask on Slack and I’ll answer later

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?

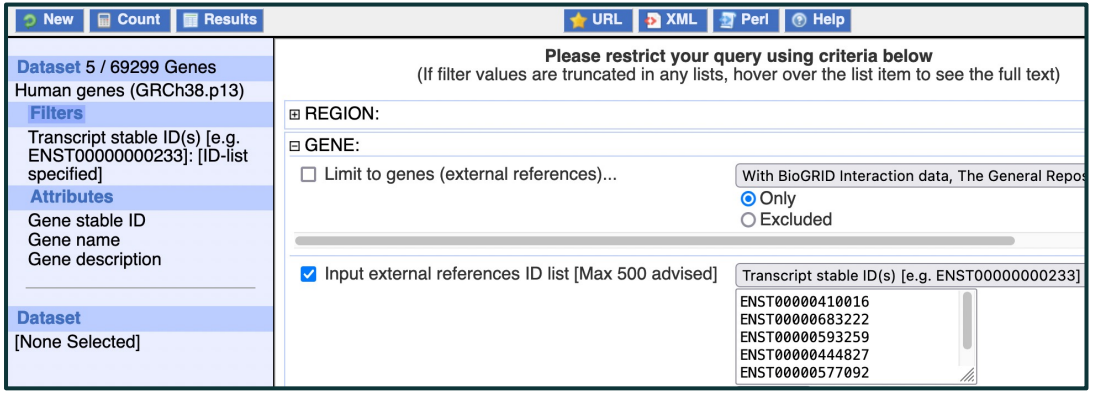
- Gene list can be:
  - Unranked (e.g. genes with somatic mutations in cancer sample)
  - Ranked (e.g. sensitivity in a CRISPR screen)
- RNA-seq differential expression analysis produces ranked lists
- Ranked lists are ordered by a score or metric:
  - e.g. adjusted p-value
  - e.g. $\log_2$ fold change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



What next?

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

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

What next?

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

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

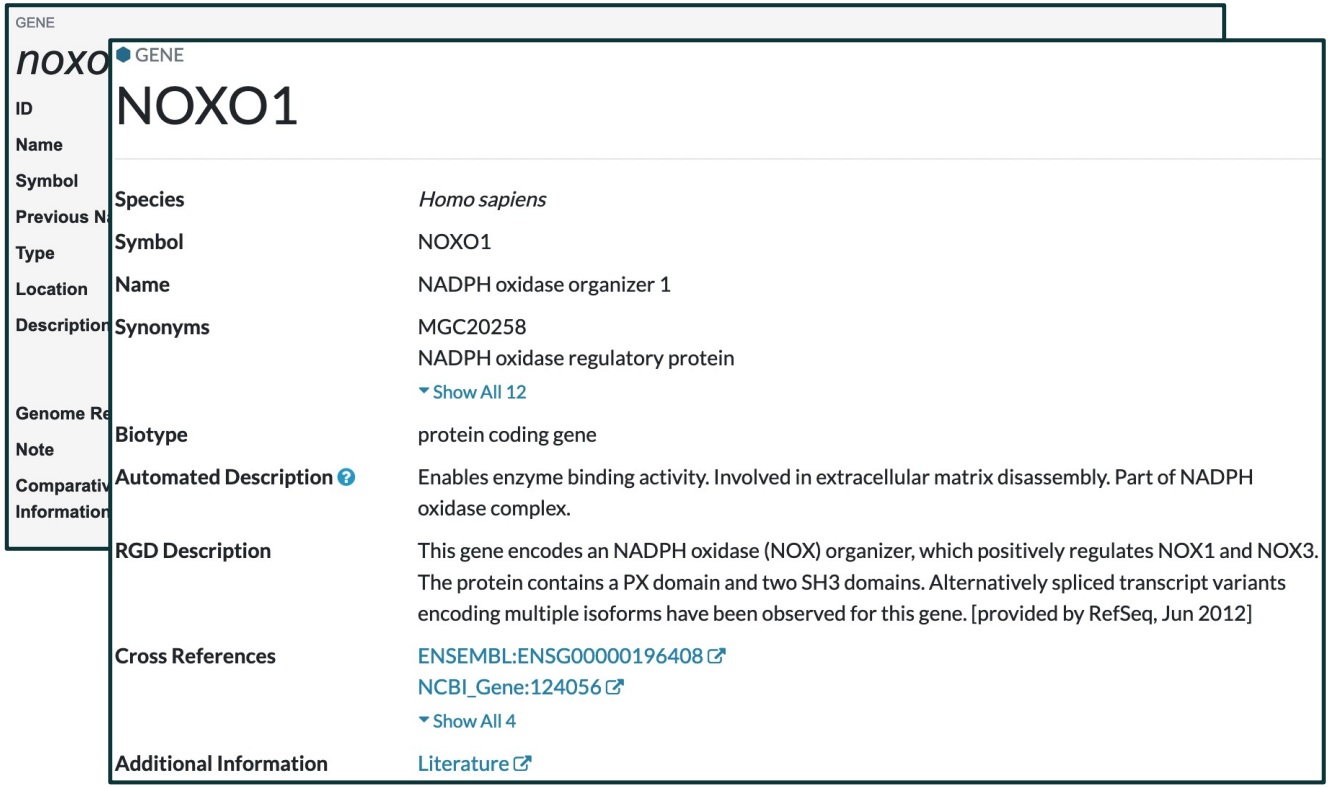
Look up genes in databases

GENE

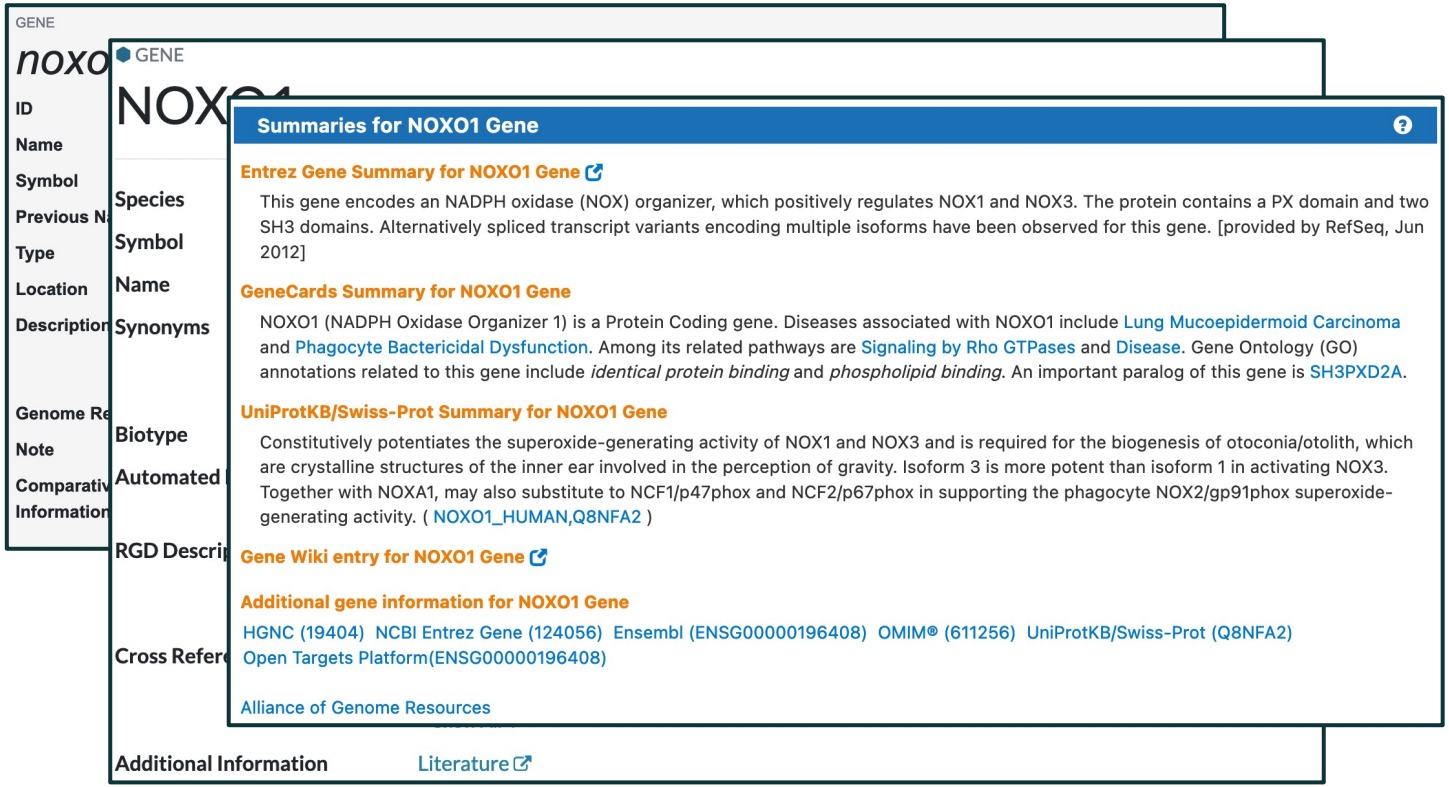
noxo1a

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	

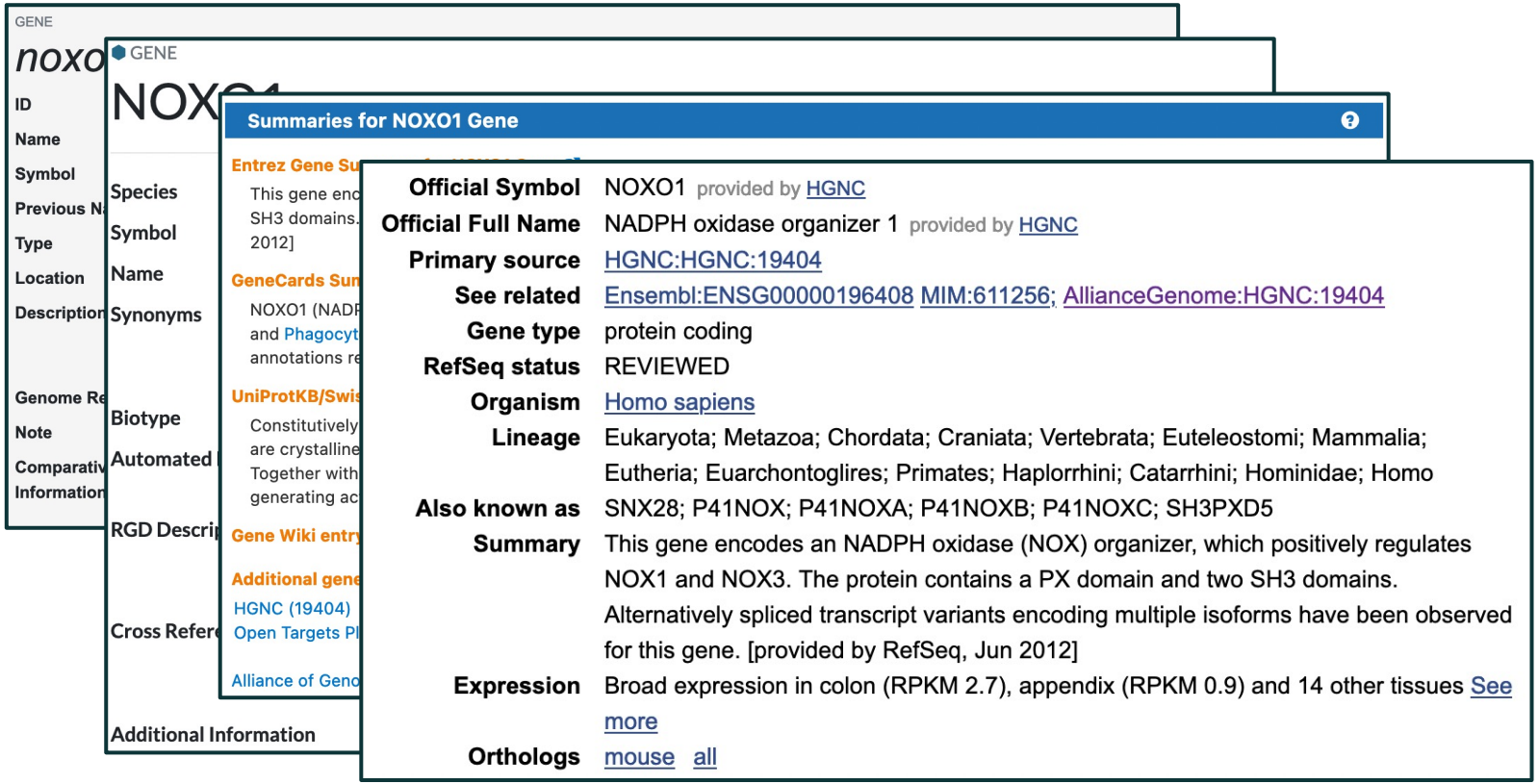
Look up genes in databases



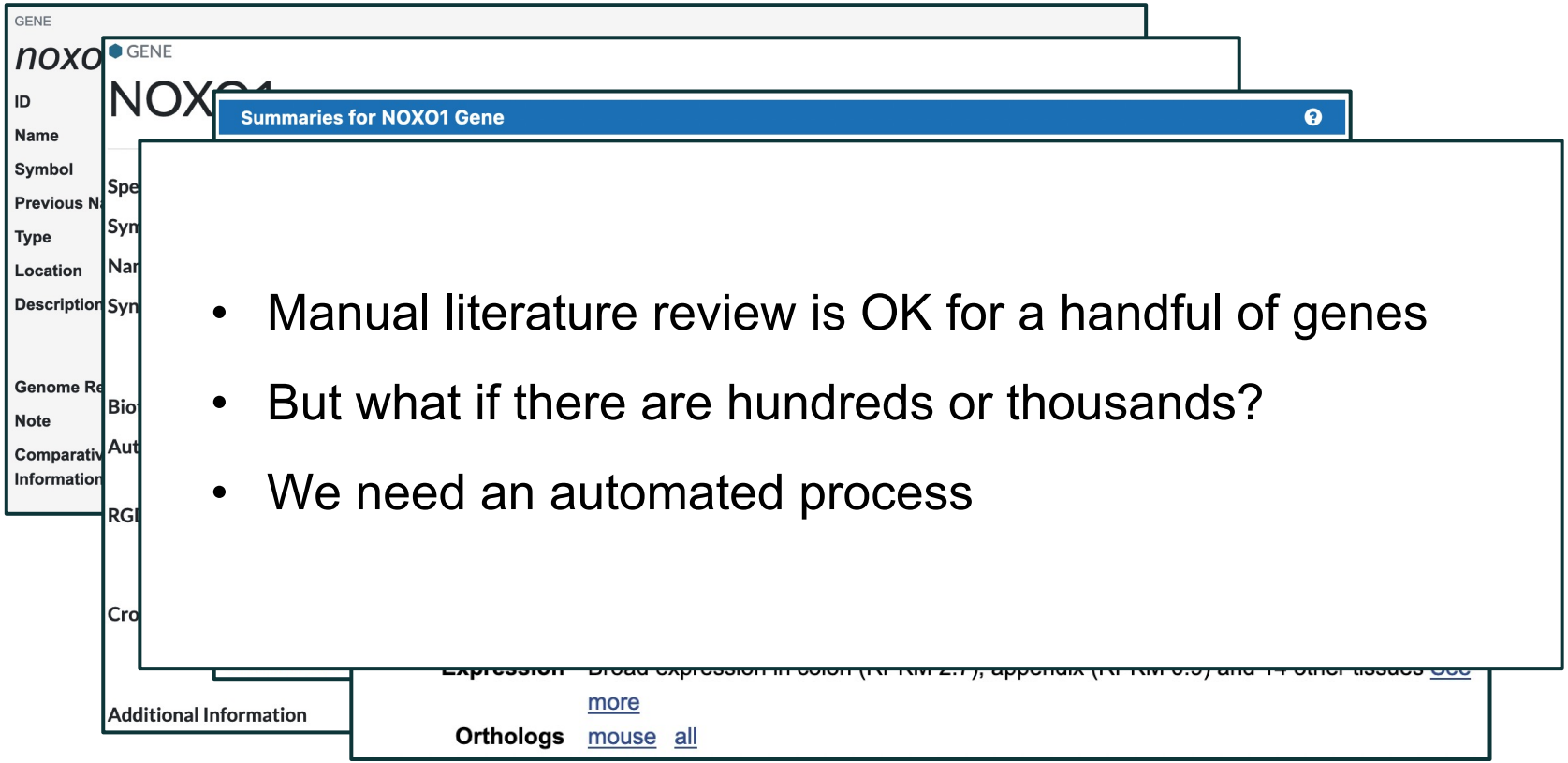
Look up genes in databases



Look up genes in databases



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 P circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	BZIP domain-containing protein
UniProtKB:A0A024QZQ1	SIRT1	involved_in	GO:0007623 P circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	Deacetylase sirtuin-type domain-containing protein
UniProtKB:A0A024R230	NTRK2	involved_in	GO:0007623 P circadian rhythm	ECO:0000265 IEA	GO_REF:0000107	Ensembl	Tyrosine-protein kinase receptor
UniProtKB:A0A024R241	NFIL3	involved_in	GO:0007623 P 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

Gene Ontology

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

GO ontologies

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

GO hierarchy

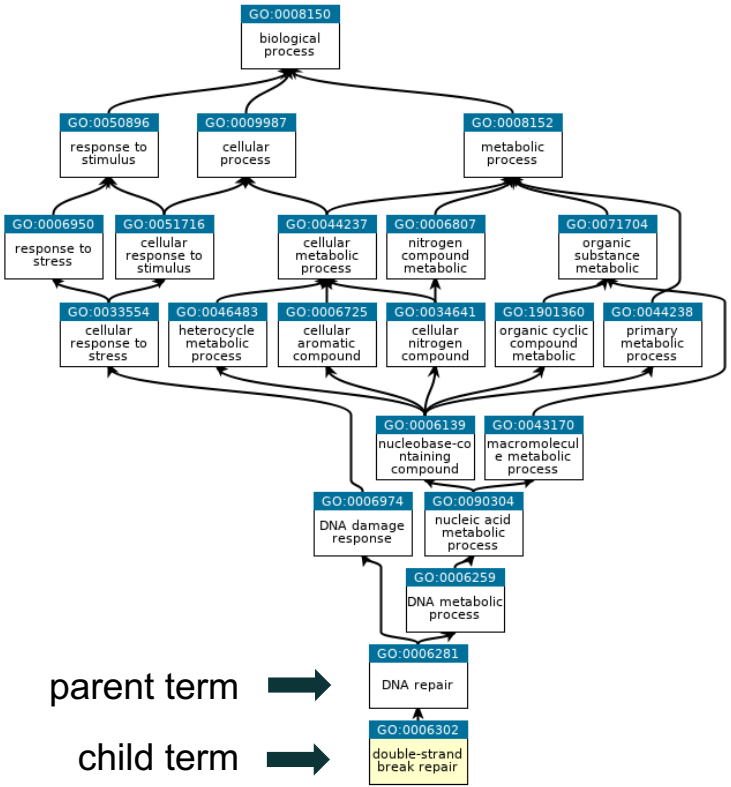
more genes

fewer genes

root term

broad terms

specific terms



BRCA2 example

Gene: BRCA2 ENSG00000139618

Description	BRCA2 DNA repair associated [Source:HGNC Symbol;Acc:HGNC:1101]
Gene Synonyms	BRCC2, FACD, FAD, FAD1, FANCD, FANCD1, XRCC11
Location	Chromosome 13: 32,315,086-32,400,268 forward strand. GRCh38:CM000675.2
About this gene	This gene has 15 transcripts (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 Show/hide columns (3 hidden) Filter

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 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

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	3.8362216674233099e-10
ENSDARG00000102758	7.675777258160306e-10
ENSDARG00000004754	1.3907777984676073e-9
ENSDARG00000099960	2.0021043563850804e-9
ENSDARG00000056615	2.0321305012048 73e-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	3.5053165649072 91e-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	5.2858127239743555e-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	2.3644634116263 39e-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	2.0021043563850804e-9
ENSDARG00000056615	2.0321305012048 73e-9
ENSDARG00000073820	3.749723988428957e-9
ENSDARG00000003570	4.0276150697004 61e-9
ENSDARG00000077799	7.4920976360234 96e-9
ENSDARG00000053836	7.754409869629003e-9
ENSDARG00000094678	7.81554132440011e-9
ENSDARG00000076914	2.8102379001185536e-8
ENSDARG00000037421	2.990985153534242e-8
ENSDARG00000014340	3.5053165649072 91e-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	5.2858127239743 55e-7
ENSDARG00000004954	5.5269704901181 69e-7
ENSDARG00000025903	7.2450677926856 57e-7
ENSDARG00000025094	7.9994748615438 12e-7
ENSDARG00000100504	8.0371254697727 79e-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	2.3644634116263 39e-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	2.0021043563850804e-9
ENSDARG00000056615	2.0321305012048 73e-9
ENSDARG00000073820	3.749723988428957e-9
ENSDARG00000003570	4.0276150697004 61e-9
ENSDARG00000077799	7.4920976360234 96e-9
ENSDARG00000053836	7.7544098696290 03e-9
ENSDARG00000094678	7.815541324400 11e-9
ENSDARG00000076914	2.8102379001185536e-8
ENSDARG00000037421	2.990985153534242e-8
ENSDARG00000014340	3.5053165649072 91e-8
ENSDARG00000104773	4.1069391754303735e-8
ENSDARG00000105749	4.1069391754303735e-8
ENSDARG00000018491	4.2756334549323 27e-8
ENSDARG00000109648	4.9384742480335793e-8
ENSDARG00000010231	5.8228595030203 41e-8
ENSDARG00000039142	7.6653816084608 26e-8
ENSDARG00000104672	1.2504789486538527e-7
ENSDARG00000054196	1.3563473037973 86e-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	4.9176672885314 57e-7
ENSDARG00000105731	5.2858127239743 55e-7
ENSDARG00000004954	5.5269704901181 69e-7
ENSDARG00000025903	7.2450677926856 57e-7
ENSDARG00000025094	7.9994748615438 12e-7
ENSDARG00000100504	8.0371254697727 79e-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	2.3644634116263 39e-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	2.0021043563850804e-9
ENSDARG00000056615	2.0321305012048 73e-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	3.5053165649072 91e-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	4.9176672885314 57e-7
ENSDARG00000105731	5.2858127239743 55e-7
ENSDARG00000004954	5.5269704901181 69e-7
ENSDARG00000025903	7.2450677926856 57e-7
ENSDARG00000025094	7.9994748615438 12e-7
ENSDARG00000100504	8.0371254697727 79e-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!

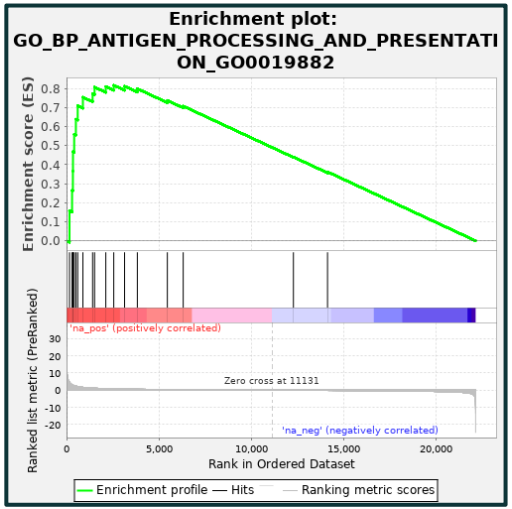
- Quiz on Mentimeter (www.menti.com)

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 1000 species/strains

Other functional enrichment tools

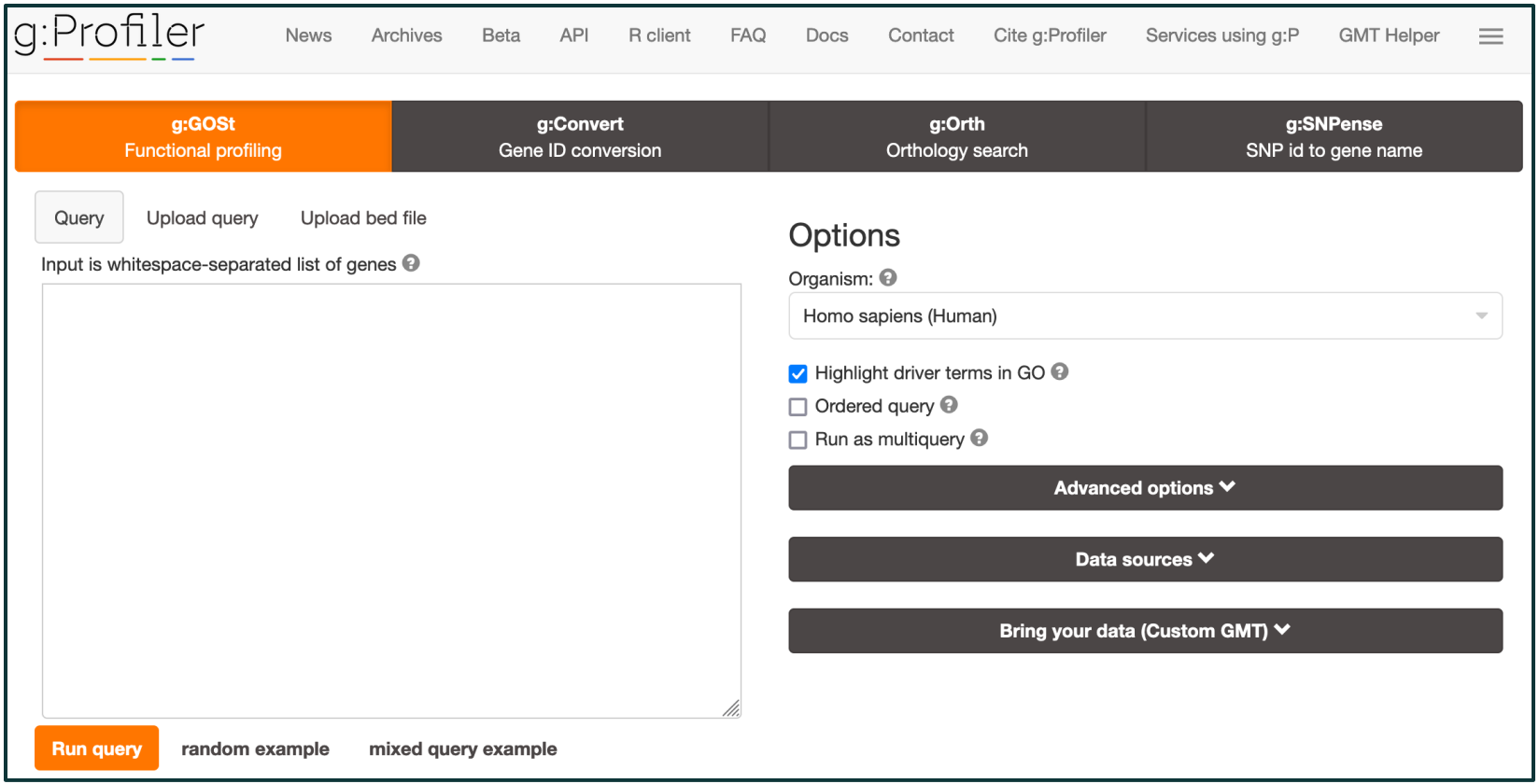
- Other tools are available (and good):
  - Enrichr (maayanlab.cloud/Enrichr/):
    - Web-based
    - Similar to g:Profiler
    - Only human, mouse, fly, yeast, worm and zebrafish
  - GSEA (www.gsea-msigdb.org/gsea/):
    - Desktop software
    - Implements GSEA method
    - Works on whole genome ranked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



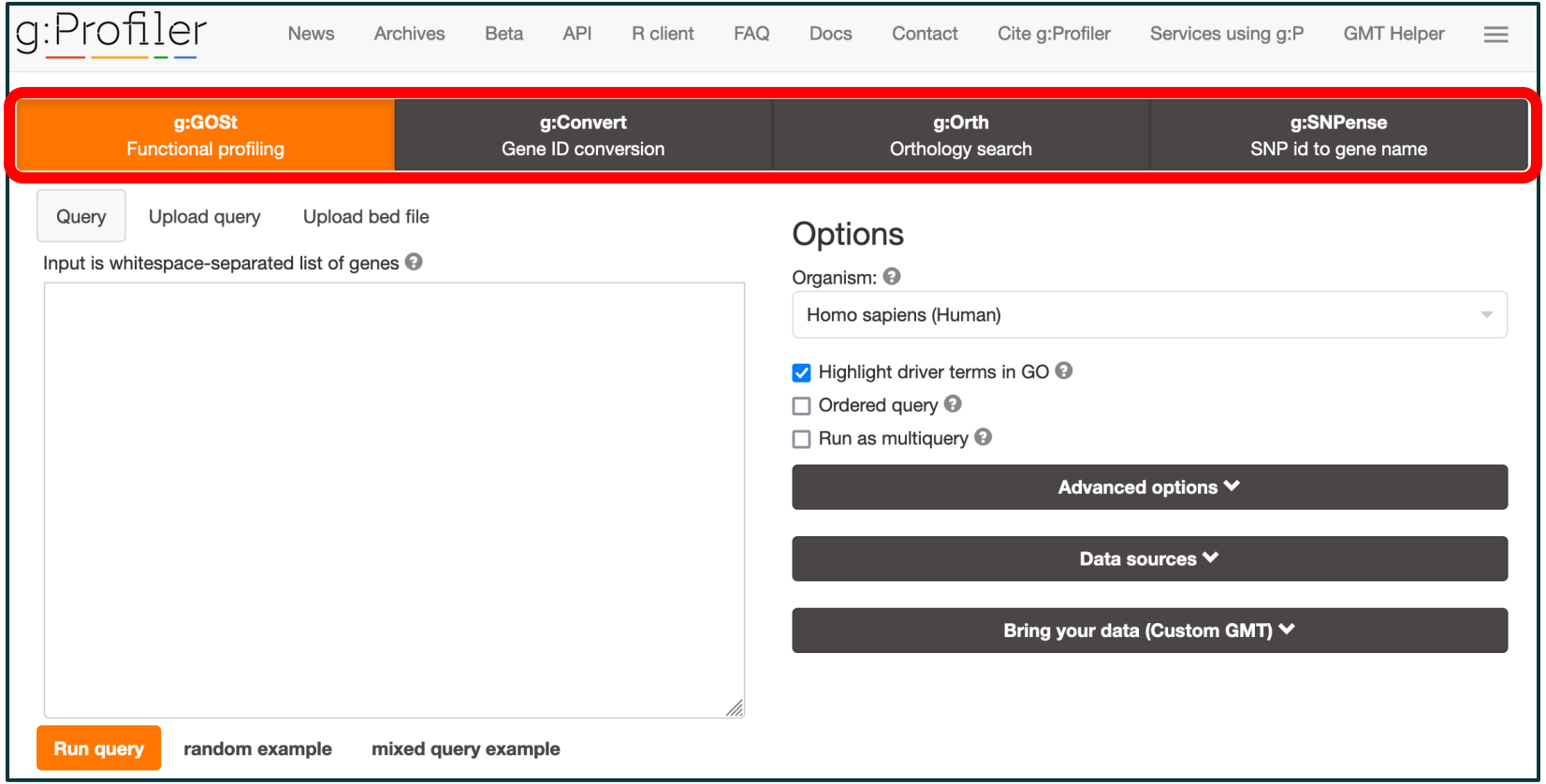
g:Profiler

- 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

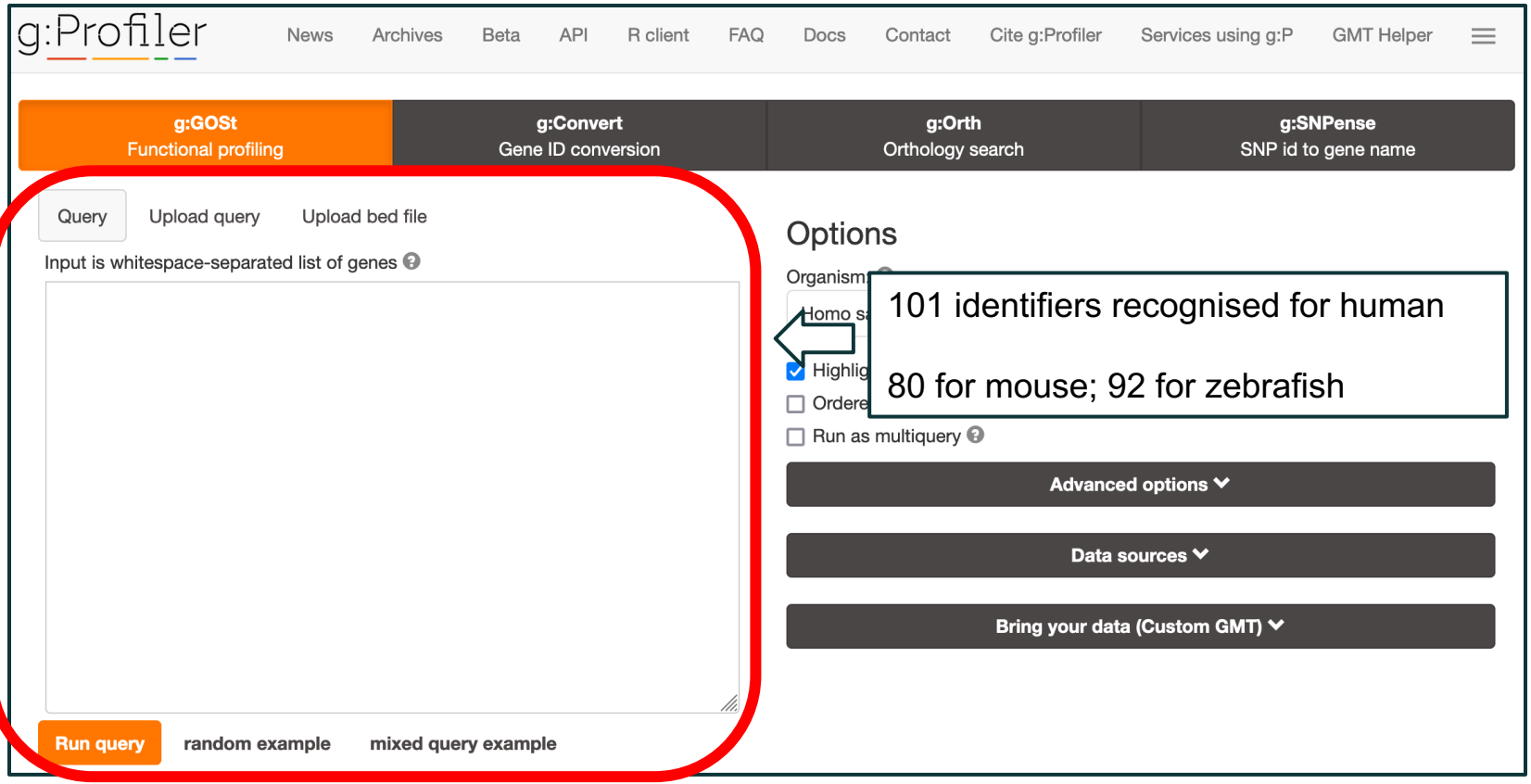
g:Profiler


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
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

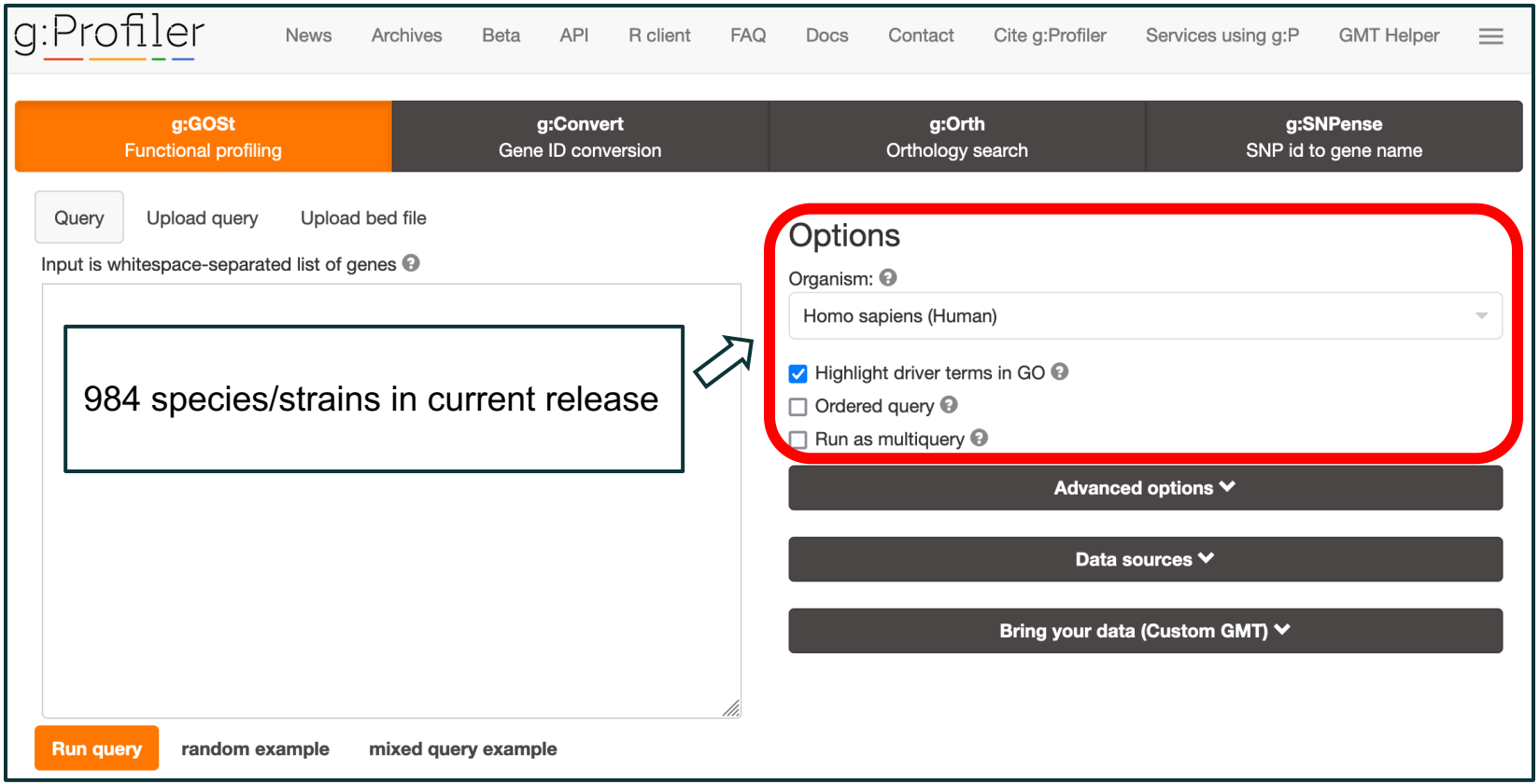
g:Profiler – four tools



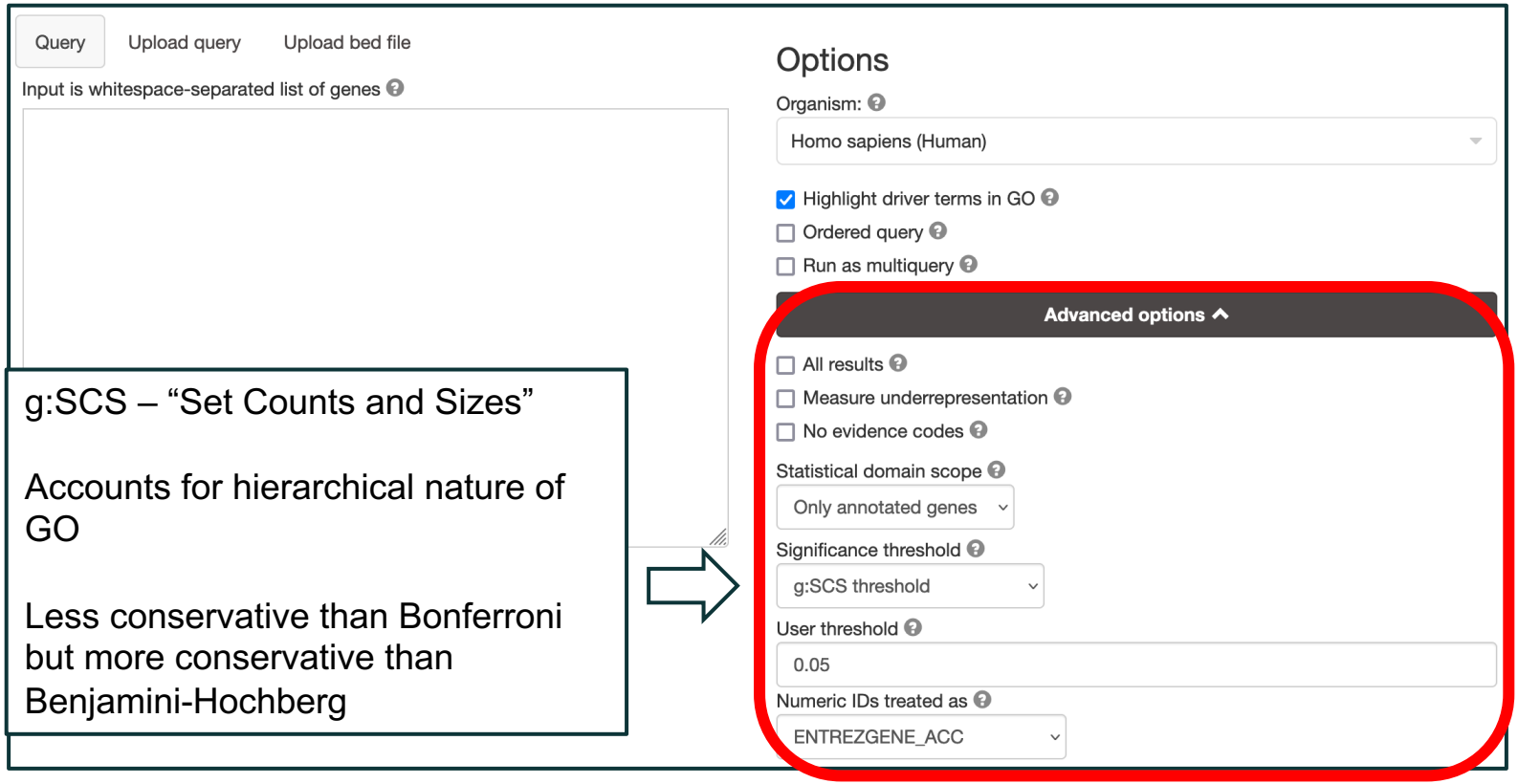
g:Profiler – gene list



g:Profiler – options



g:Profiler – advanced options

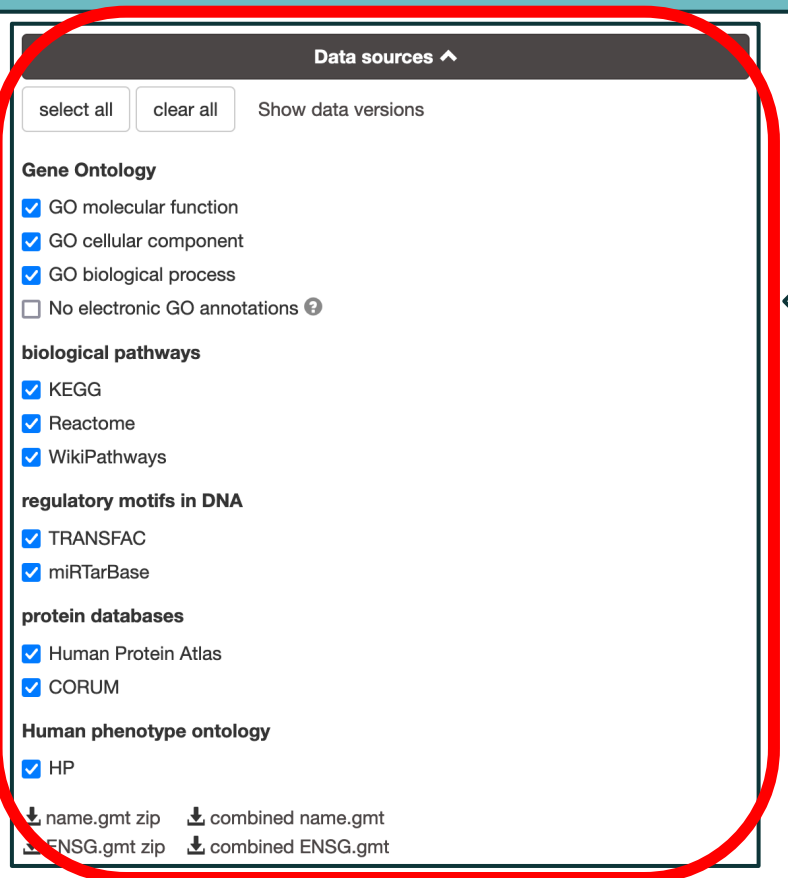


g:Profiler – data sources

9 data sources

(or 11 if count GO as three separate sources)

All 9 not available for all species

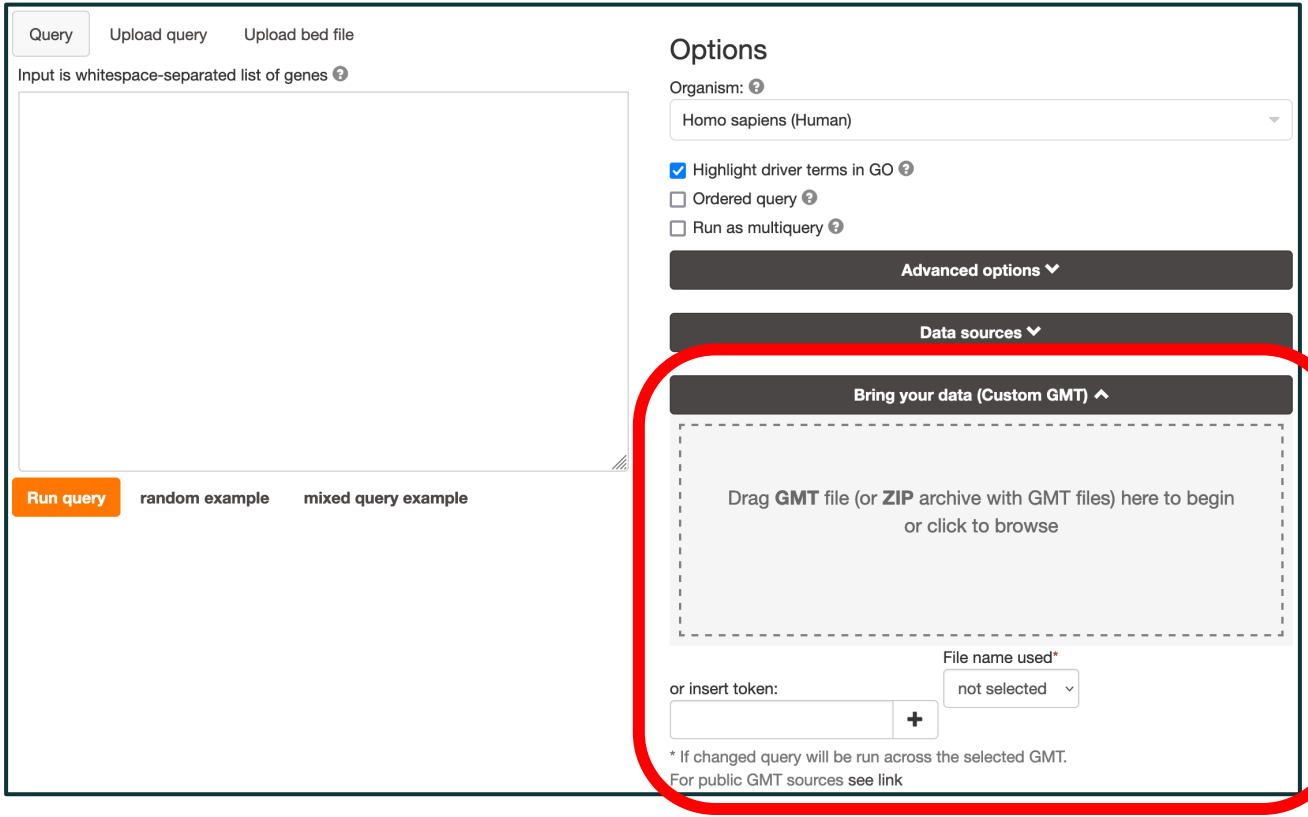


Can exclude GO IEA evidence term (inferred from electronic annotation)

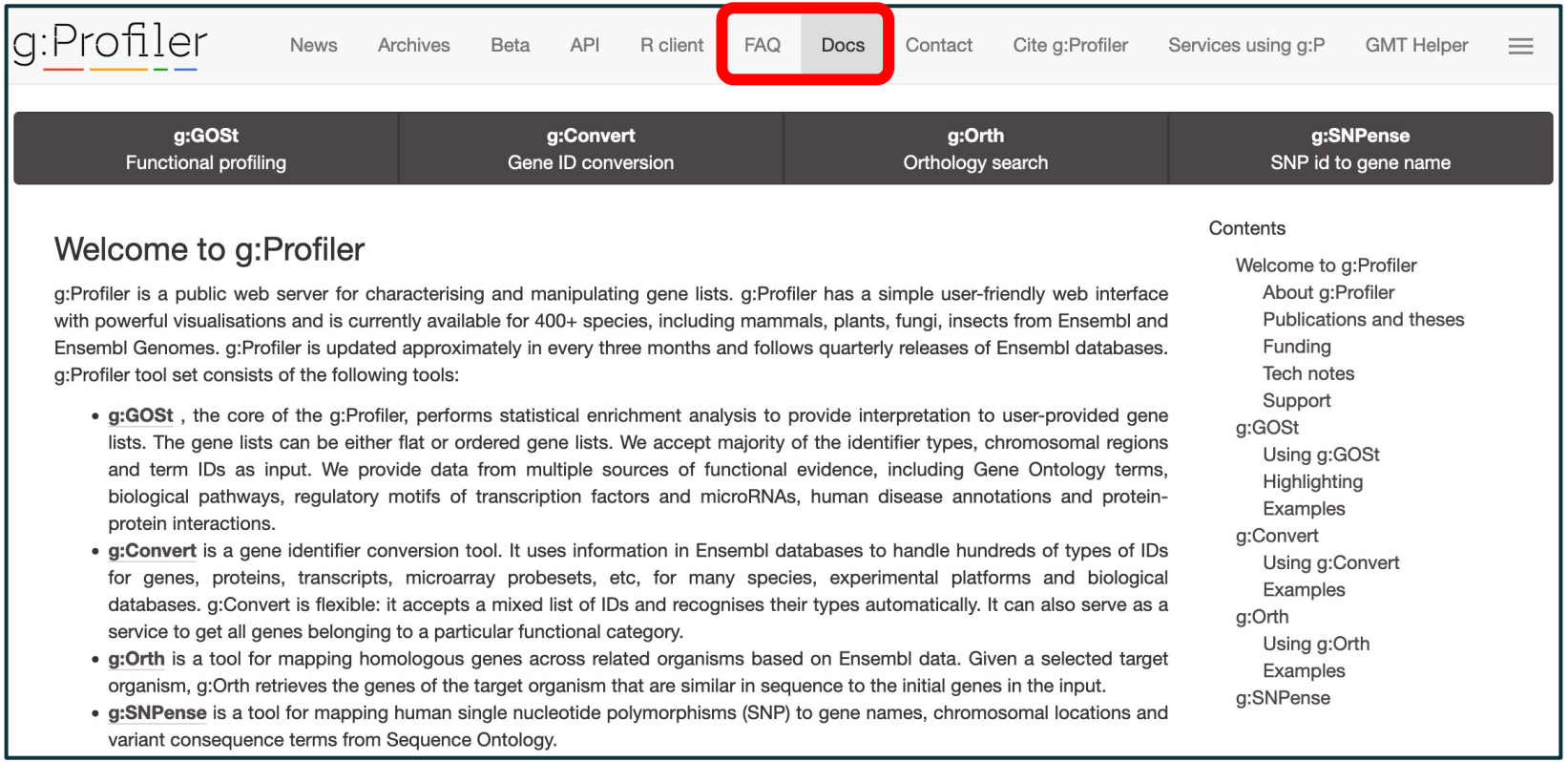
But often as reliable as human annotation (Škunca *et al.* 2012)

Suggest running with and without if using human or model organisms

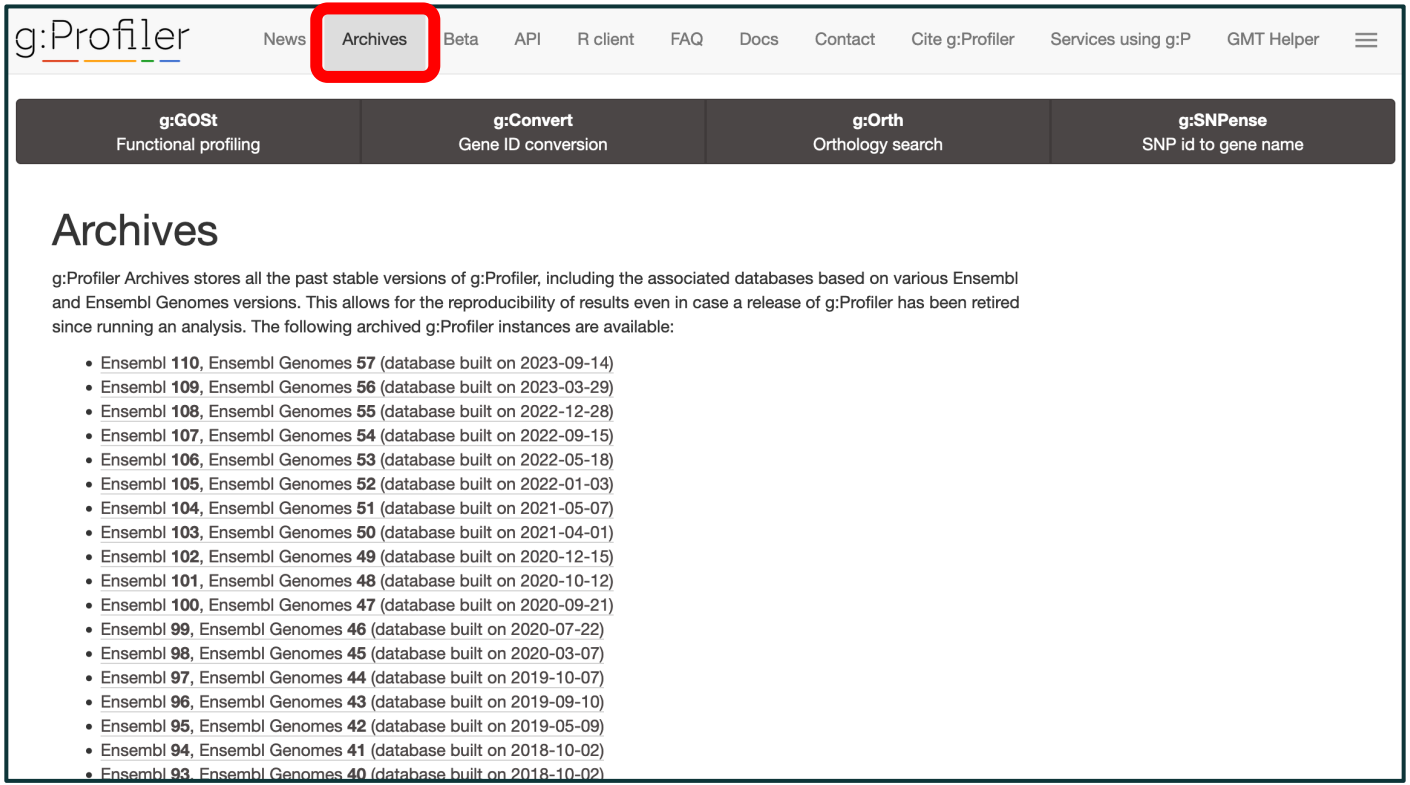
g:Profiler – bring your data



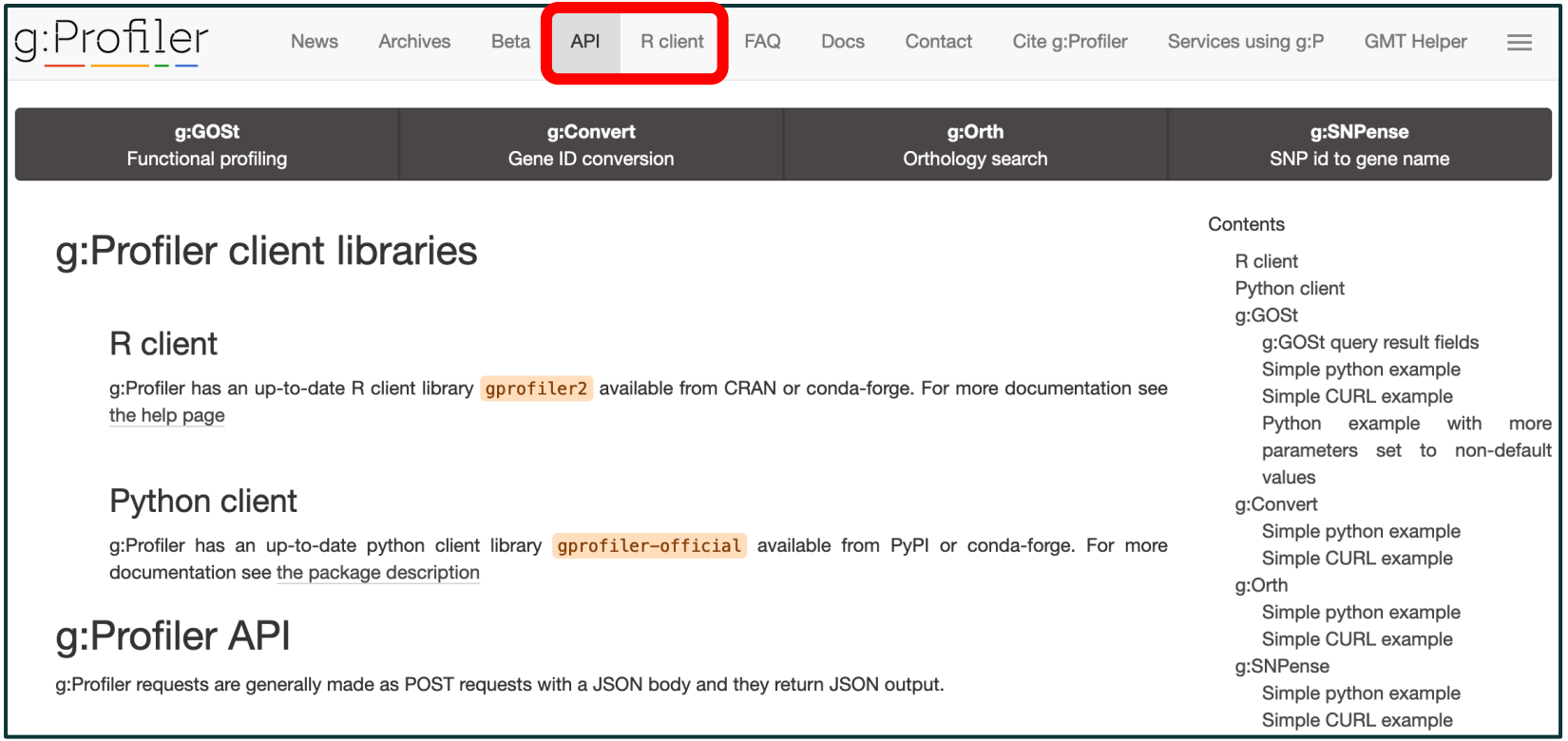
g:Profiler – documentation



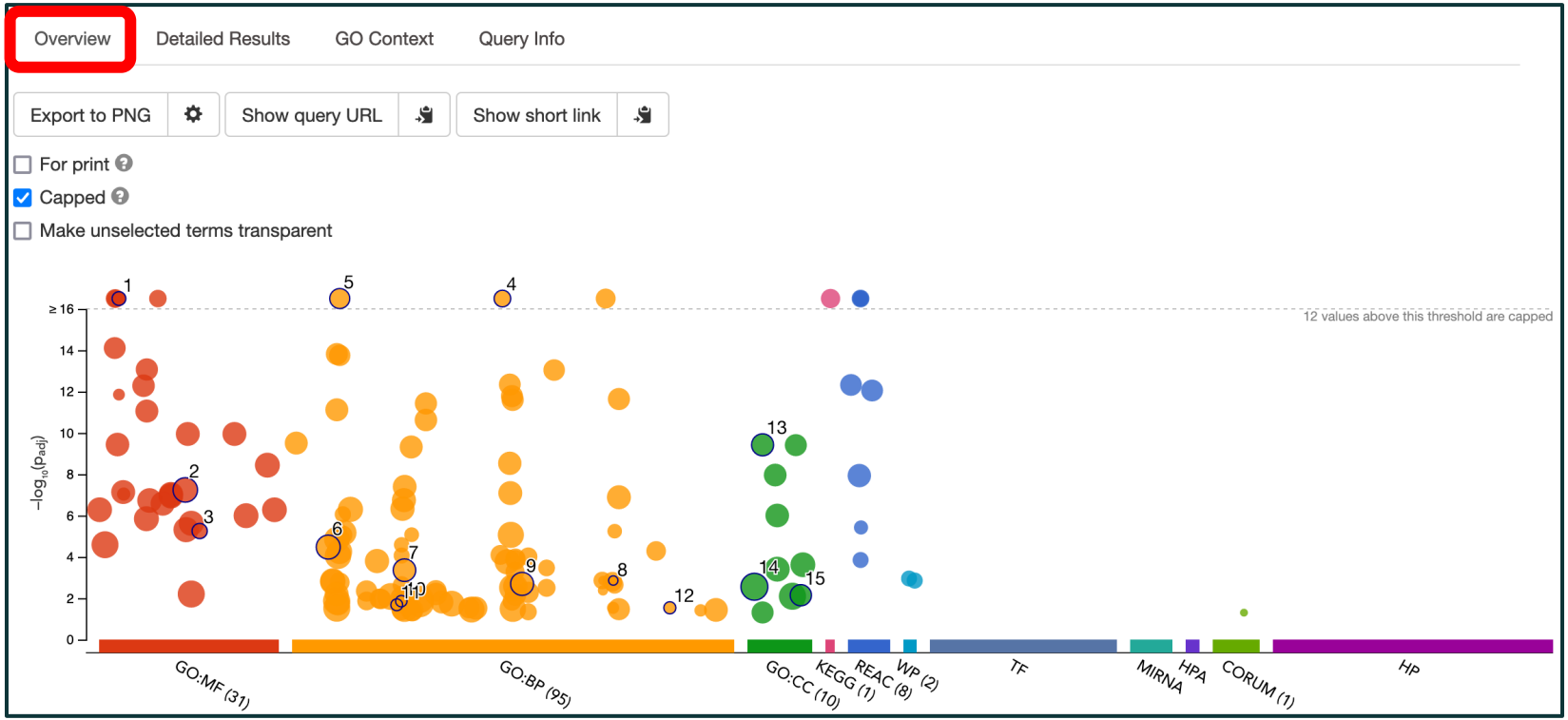
g:Profiler – archives



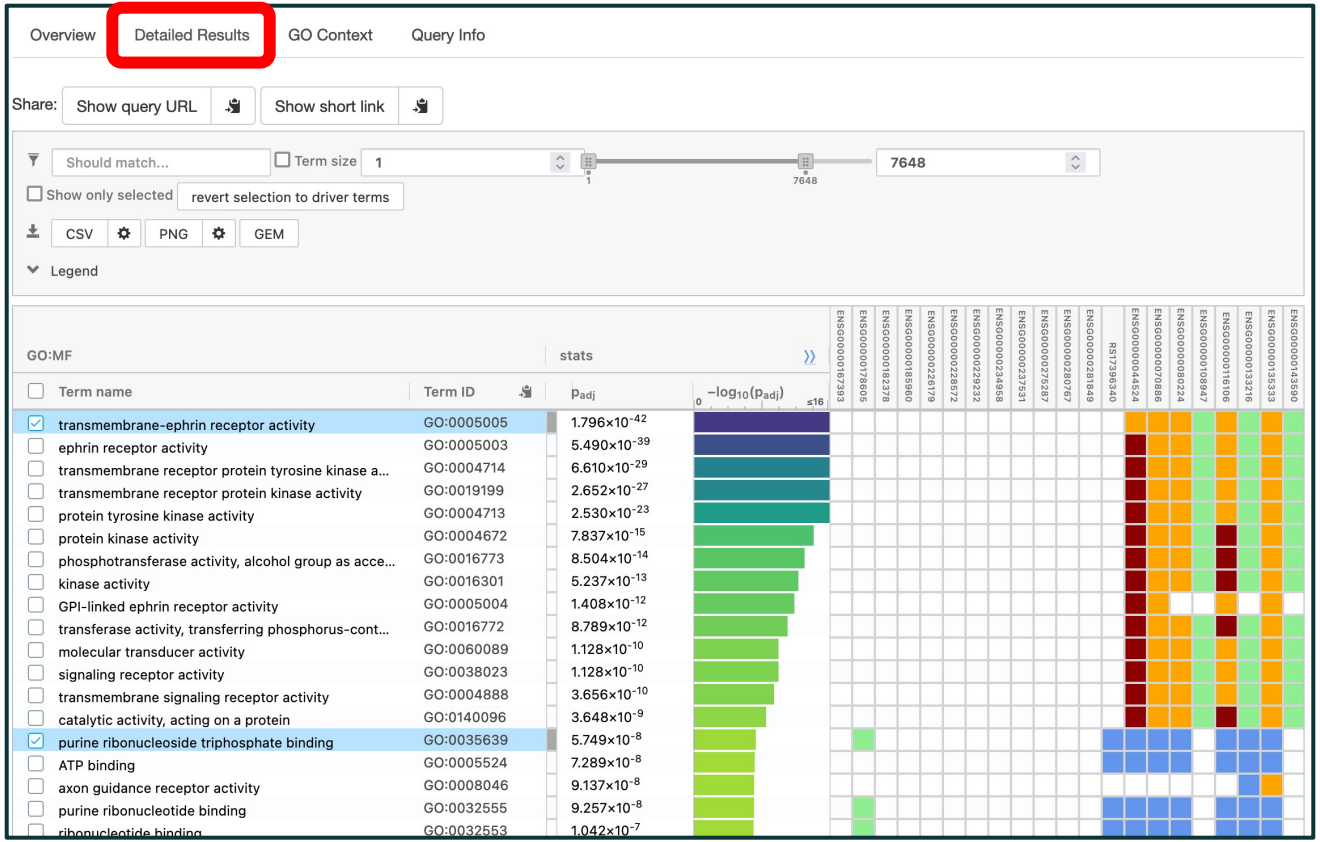
g:Profiler – API and libraries



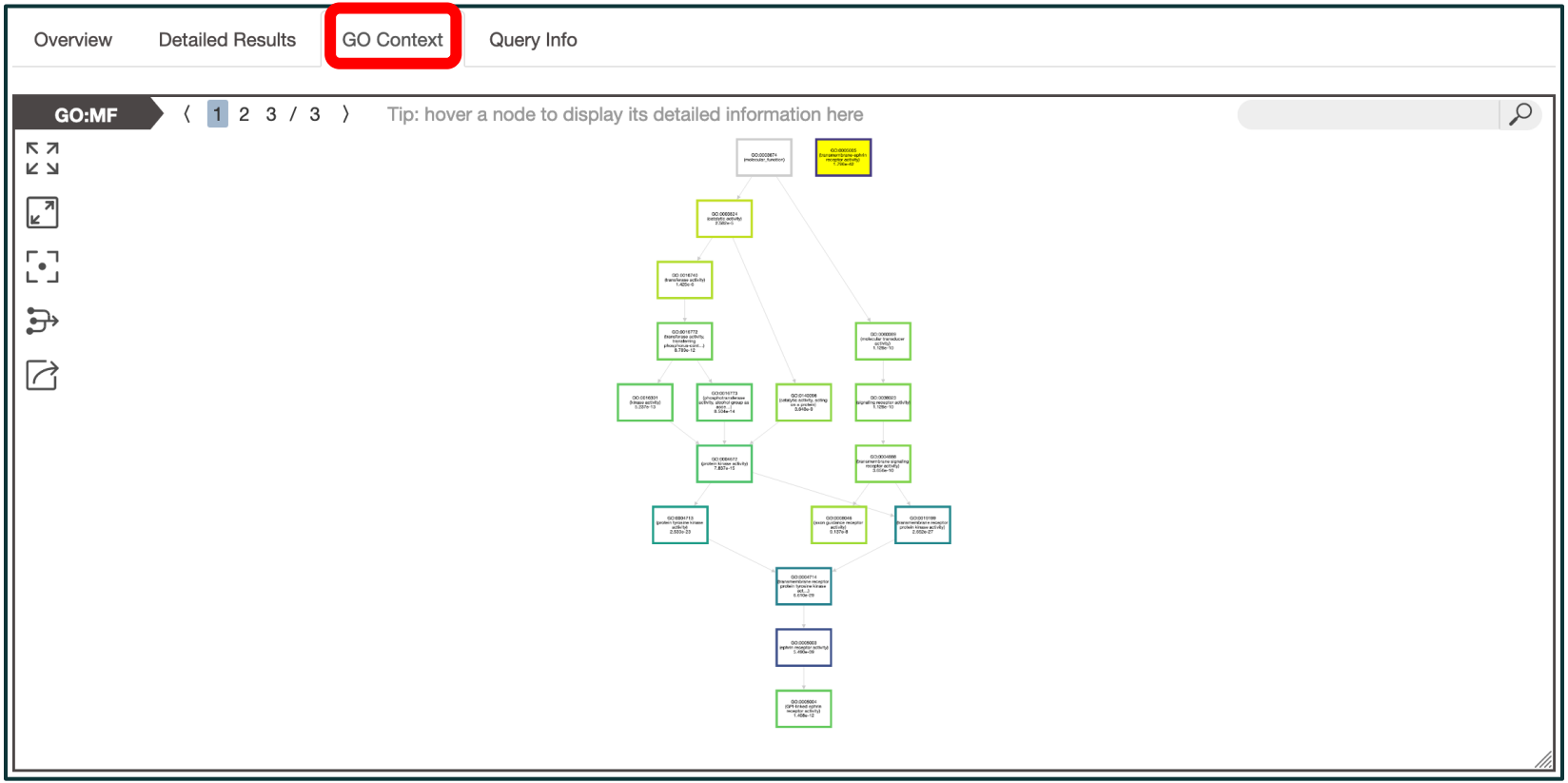
g:Profiler – overview



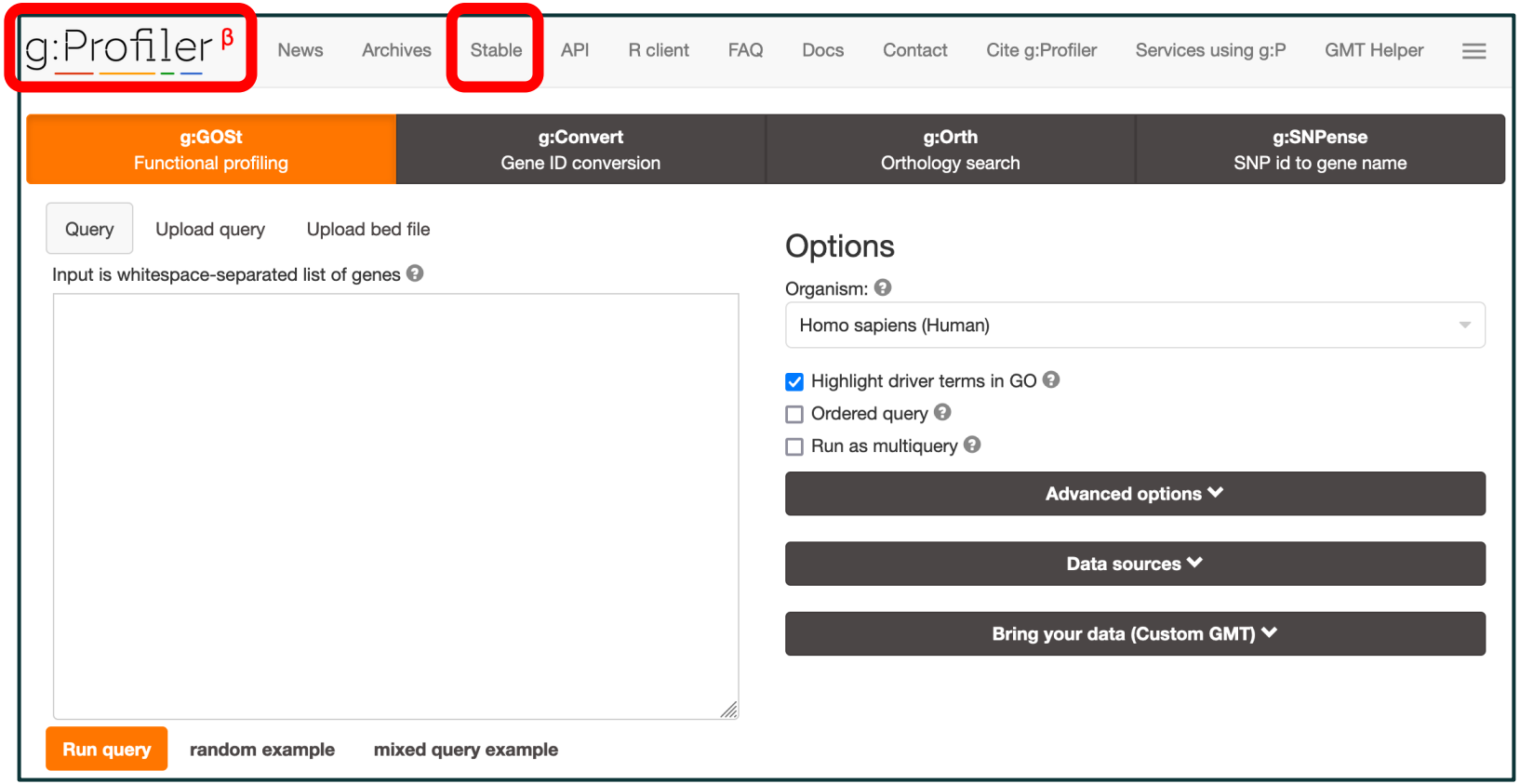
g:Profiler – detailed results



g:Profiler – GO context

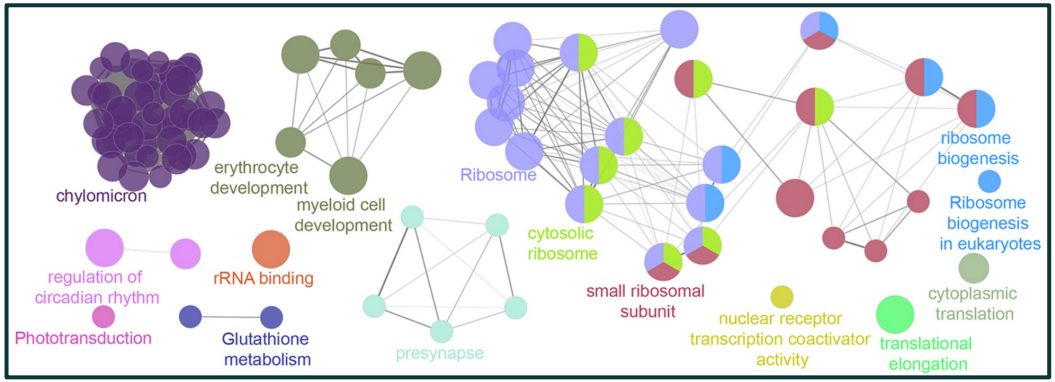


g:Profiler – beta



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