GO analysis walkthrough

By Mary Allen, Jesse Kurland

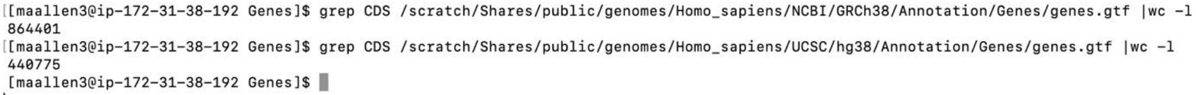
How does GO term enrichment work?

- scRNA-seq on mouse skeletal muscle
  - Compare to mm10 genome? -> “muscle”
  - Compare to all genes expressed in dataset? -> Identifies different myogenic populations
- Example:
  - Aged vs Adult sRNA-seq from mouse muscle -> 1000 differentially expressed genes in Aged mice
    - **In background gene set**
      - 100,000 total genes in mm10
      - 100 genes involved in innervation of skeletal muscle
    - **In differentially expressed gene set**
      - 100 genes involved in innervation → not significant!
      - 200 genes involved in innervation → **significant**!

Gathering the gene lists

Before you run Deseq2

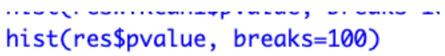
Decide on which GTF you will use because some gtfs have more genes than others



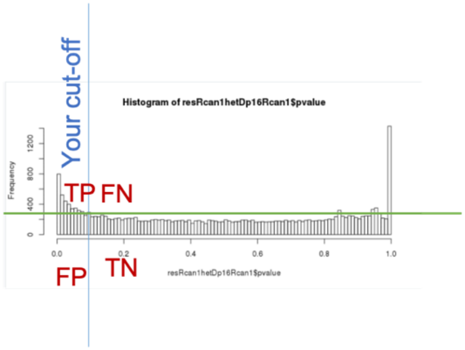
```
[maallen3@ip-172-31-38-192 Genes]$ grep CDS /scratch/Shares/public/genomes/Homo_sapiens/NCBI/GRCh38/Annotation/Genes/genes.gtf |wc -l
864401
[maallen3@ip-172-31-38-192 Genes]$ grep CDS /scratch/Shares/public/genomes/Homo_sapiens/UCSC/hg38/Annotation/Genes/genes.gtf |wc -l
440775
[maallen3@ip-172-31-38-192 Genes]$ 
```

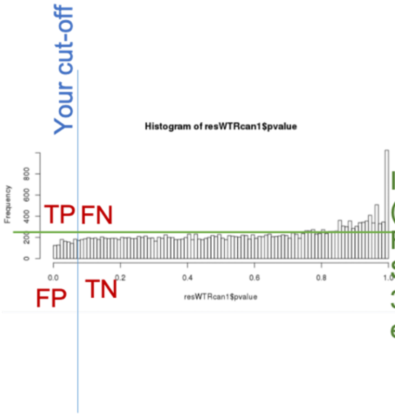
Pro for NCBI/Ensable gtfs: they have way more non-coding RNAs
Con for NCBI/Ensable gtfs: they have way more non-coding RNAs, which means more multiple hypothesis correction and therefore less significant differentaily expressed genes.

How do I pick my Deseq2 cuttoff?



```
hist(res$pvalue, breaks=100)
```





If every bin was same
(i.e. nothing was changing)
For example, 6 replicates of the
Same thing called 3 reps of A and
3 reps of bed you should have
even levels in every bin.

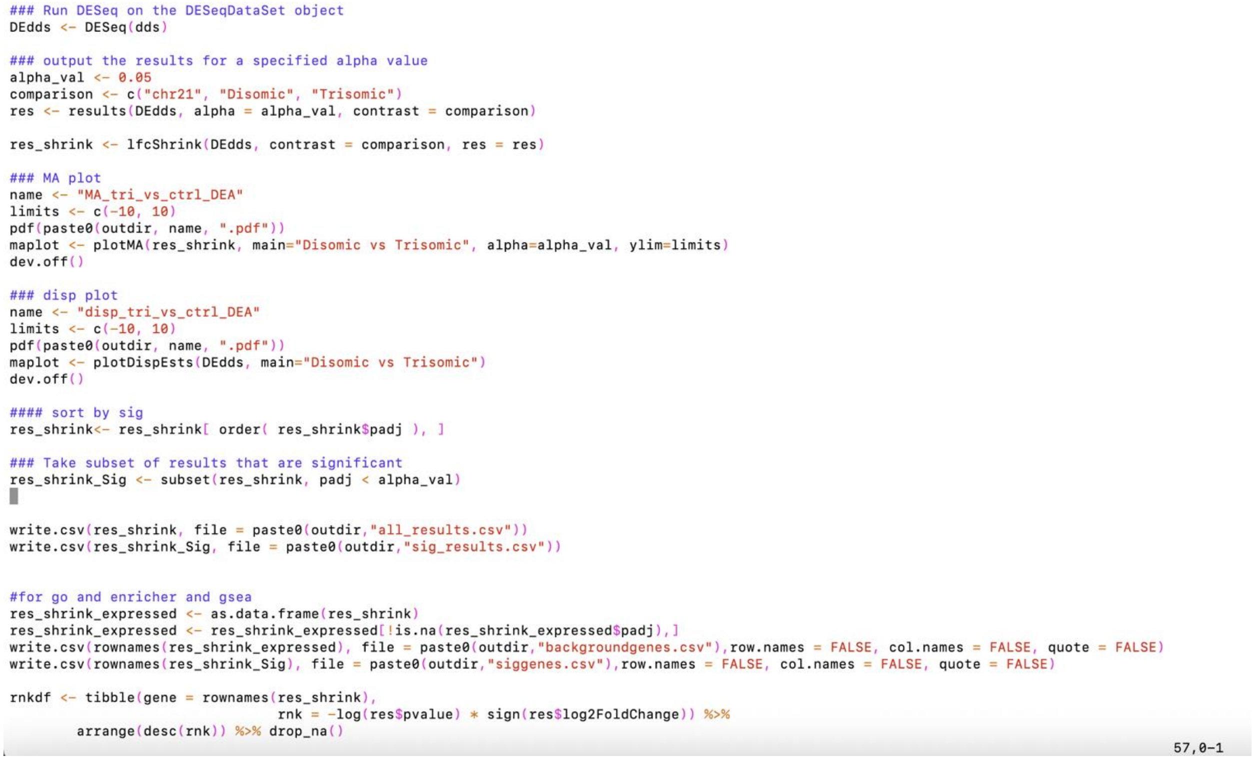
Draw a histogram of your res$pvalue
Image a blue line at your cut-off and a green line that goes flat across the bins.
These two lines help you to think about your True Positives, False Postives, True Negatives, and False negatives. If you reduce your cut off you get less genes as significant, but more of them are true positives and less of them are false positives.

How do get my gene lists out of R from Deseq2?

To run GO you will need a significantly different genes list and a background gene list.

- Background gene lists? Which one?
  If you could not have called it as differentially expressed it should not be in your background gene list.

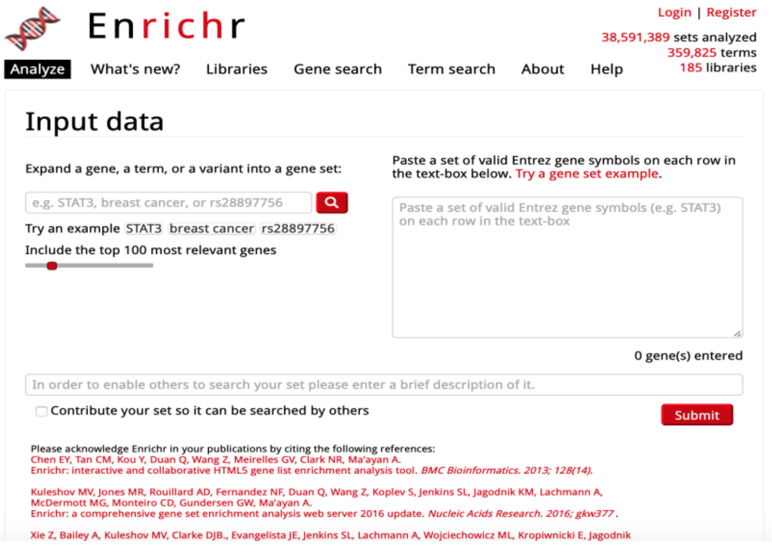
The last few lines of this script gather your background gene list and you significant gene list. Genes that are two low or variable to test for differential expression get a NA in the padj column.



```
### Run DESeq on the DESeqDataSet object
DEdds <- DESeq(dds)

### output the results for a specified alpha value
alpha_val <- 0.05
comparison <- c("chr21", "Disomic", "Trisomic")
res <- results(DEdds, alpha = alpha_val, contrast = comparison)

res_shrink <- lfcShrink(DEdds, contrast = comparison, res = res)

### MA plot
name <- "MA_tri_vs_ctrl_DEA"
limits <- c(-10, 10)
pdf(paste0(outdir, name, ".pdf"))
maplot <- plotMA(res_shrink, main="Disomic vs Trisomic", alpha=alpha_val, ylim=limits)
dev.off()

### disp plot
name <- "disp_tri_vs_ctrl_DEA"
limits <- c(-10, 10)
pdf(paste0(outdir, name, ".pdf"))
maplot <- plotDispEsts(DEdds, main="Disomic vs Trisomic")
dev.off()

#### sort by sig
res_shrink<- res_shrink[ order( res_shrink$padj ), ]

### Take subset of results that are significant
res_shrink_Sig <- subset(res_shrink, padj < alpha_val)


write.csv(res_shrink, file = paste0(outdir,"all_results.csv"))
write.csv(res_shrink_Sig, file = paste0(outdir,"sig_results.csv"))


#for go and enricher and gsea
res_shrink_expressed <- as.data.frame(res_shrink)
res_shrink_expressed <- res_shrink_expressed[!is.na(res_shrink_expressed$padj),]
write.csv(rownames(res_shrink_expressed), file = paste0(outdir,"backgroundgenes.csv"),row.names = FALSE, col.names = FALSE, quote = FALSE)
write.csv(rownames(res_shrink_Sig), file = paste0(outdir,"siggenes.csv"),row.names = FALSE, col.names = FALSE, quote = FALSE)

rnkdf <- tibble(gene = rownames(res_shrink),
                                rnk = -log(res$pvalue) * sign(res$log2FoldChange)) %>%
        arrange(desc(rnk)) %>% drop_na()
                                                                                                    57,0-1
```

Enrichr (maayanlab.cloud/Enrichr/)

- Conducts multiple comparisons (doesn't permit using background gene set)
- Great for first pass checks of what you should explore more… not the most statically accurate (since not using real background lists)



Input data

Expand a gene, a term, or a variant into a gene set:

e.g. STAT3, breast cancer, or rs28897756

Try an example STAT3 breast cancer rs28897756

Include the top 100 most relevant genes

Paste a set of valid Entrez gene symbols on each row in the text-box below. Try a gene set example.

Paste a set of valid Entrez gene symbols (e.g. STAT3) on each row in the text-box

0 gene(s) entered

In order to enable others to search your set please enter a brief description of it.

□ Contribute your set so it can be searched by others

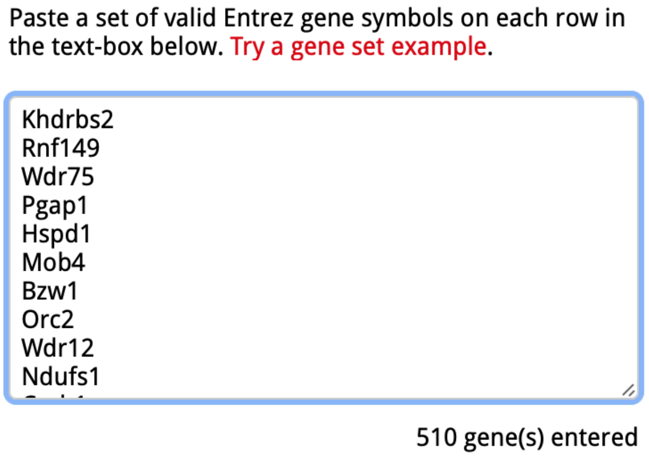
Submit

Please acknowledge Enrichr in your publications by citing the following references:
Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma'ayan A.
Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics. 2013; 128(14).*

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, McDermott MG, Monteiro CD, Gundersen GW, Ma'ayan A.
Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research. 2016; gkw377* .

Xie Z, Bailey A, Kuleshov MV, Clarke DJB., Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik

- Paste enriched gene list into box and “submit”

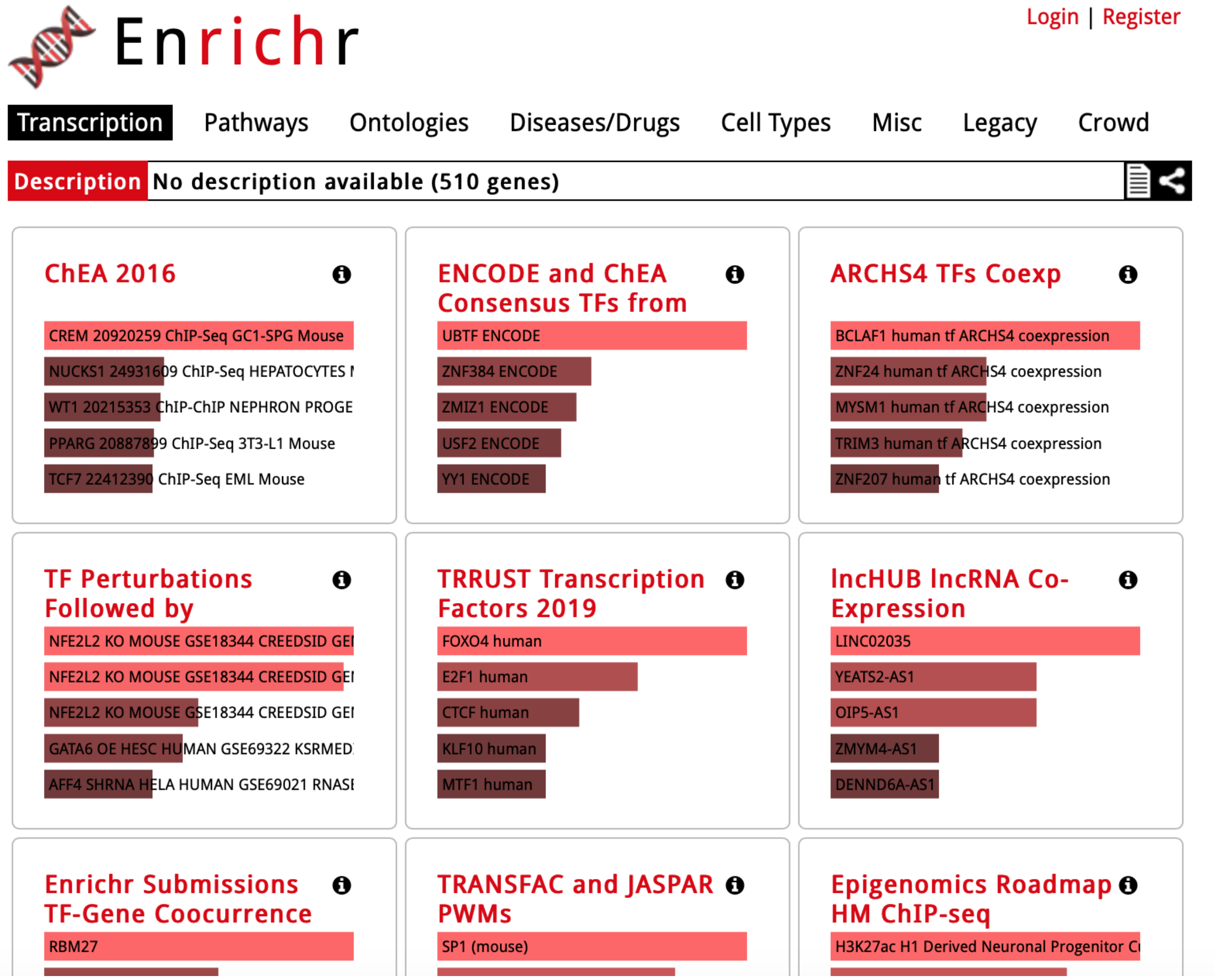


Paste a set of valid Entrez gene symbols on each row in the text-box below. Try a gene set example.

```
Khdrbs2
Rnf149
Wdr75
Pgap1
Hspd1
Mob4
Bzw1
Orc2
Wdr12
Ndufs1
```

510 gene(s) entered

- “submit”



Description No description available (510 genes)

ChEA 2016
CREM 20920259 ChIP-Seq GC1-SPG Mouse
NUCKS1 24931609 ChIP-Seq HEPATOCYTES
WT1 20215353 ChIP-ChIP NEPHRON PROGE
PPARG 20887899 ChIP-Seq 3T3-L1 Mouse
TCF7 22412390 ChIP-Seq EML Mouse

ENCODE and ChEA Consensus TFs from
UBTF ENCODE
ZNF384 ENCODE
ZMIZ1 ENCODE
USF2 ENCODE
YY1 ENCODE

ARCHS4 TFs Coexp
BCLAF1 human tf ARCHS4 coexpression
ZNF24 human tf ARCHS4 coexpression
MYSM1 human tf ARCHS4 coexpression
TRIM3 human tf ARCHS4 coexpression
ZNF207 human tf ARCHS4 coexpression

TF Perturbations Followed by
NFE2L2 KO MOUSE GSE18344 CREEDSID GE
NFE2L2 KO MOUSE GSE18344 CREEDSID GE
NFE2L2 KO MOUSE GSE18344 CREEDSID GE
GATA6 OE HESC HUMAN GSE69322 KSRMED
AFF4 SHRNA HELA HUMAN GSE69021 RNASE

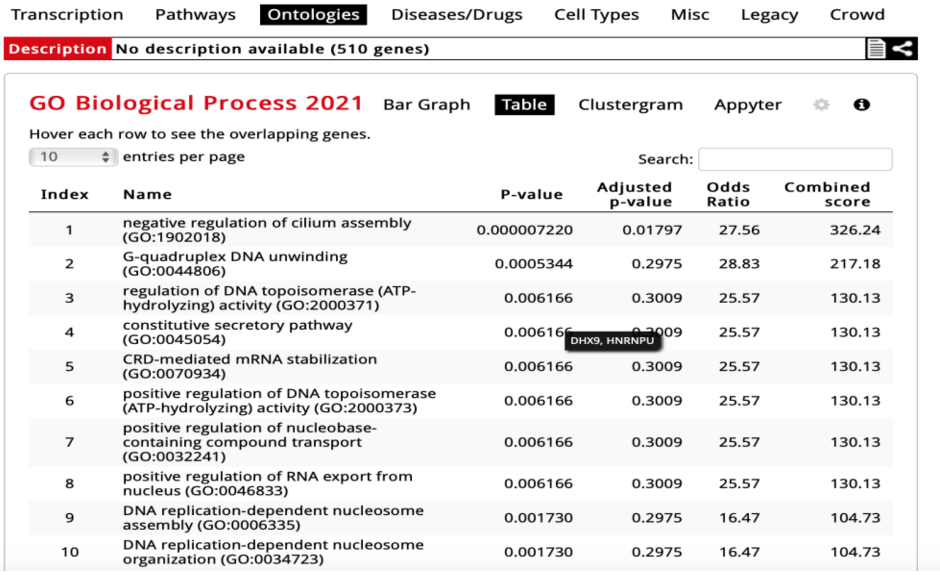
TRRUST Transcription Factors 2019
FOXO4 human
E2F1 human
CTCF human
KLF10 human
MTF1 human

lncHUB lncRNA Co-Expression
LINC02035
YEATS2-AS1
OIP5-AS1
ZMYM4-AS1
DENND6A-AS1

Enrichr Submissions TF-Gene Coocurrence
RBM27

TRANSFAC and JASPAR PWMs
SP1 (mouse)

Epigenomics Roadmap HM ChIP-seq
H3K27ac H1 Derived Neuronal Progenitor C



Description No description available (510 genes)

GO Biological Process 2021

Hover each row to see the overlapping genes.

10 entries per page

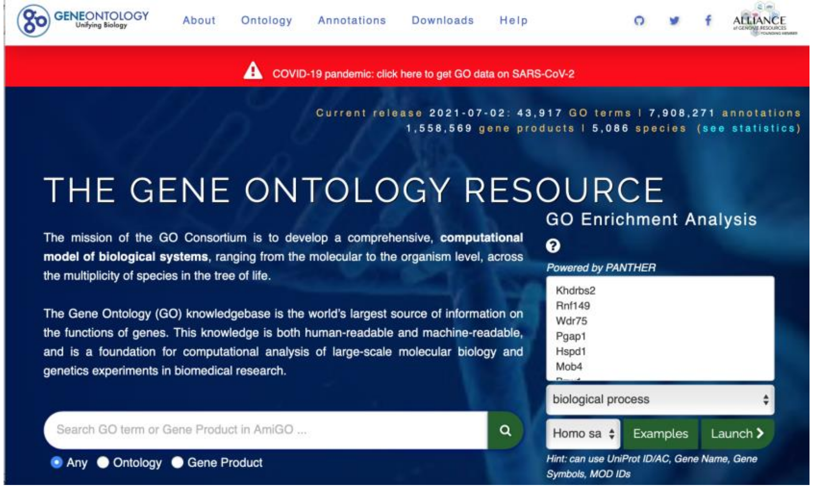
Search:

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	negative regulation of cilium assembly (GO:1902018)	0.000007220	0.01797	27.56	326.24
2	G-quadruplex DNA unwinding (GO:0044806)	0.0005344	0.2975	28.83	217.18
3	regulation of DNA topoisomerase (ATP-hydrolyzing) activity (GO:2000371)	0.006166	0.3009	25.57	130.13
4	constitutive secretory pathway (GO:0045054)	0.006166	0.3009	25.57	130.13
5	CRD-mediated mRNA stabilization (GO:0070934)	0.006166	0.3009	25.57	130.13
6	positive regulation of DNA topoisomerase (ATP-hydrolyzing) activity (GO:2000373)	0.006166	0.3009	25.57	130.13
7	positive regulation of nucleobase-containing compound transport (GO:0032241)	0.006166	0.3009	25.57	130.13
8	positive regulation of RNA export from nucleus (GO:0046833)	0.006166	0.3009	25.57	130.13
9	DNA replication-dependent nucleosome assembly (GO:0006335)	0.001730	0.2975	16.47	104.73
10	DNA replication-dependent nucleosome organization (GO:0034723)	0.001730	0.2975	16.47	104.73

DHX9, HNRNPU

Panther (http://geneontology.org)

- Allows using background gene sets
- Provides “Molecular Pathways”



COVID-19 pandemic: click here to get GO data on SARS-CoV-2

Current release 2021-07-02: 43,917 GO terms | 7,908,271 annotations
1,558,569 gene products | 5,086 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.

Search GO term or Gene Product in AmiGO ...

● Any ○ Ontology ○ Gene Product

GO Enrichment Analysis

Powered by PANTHER

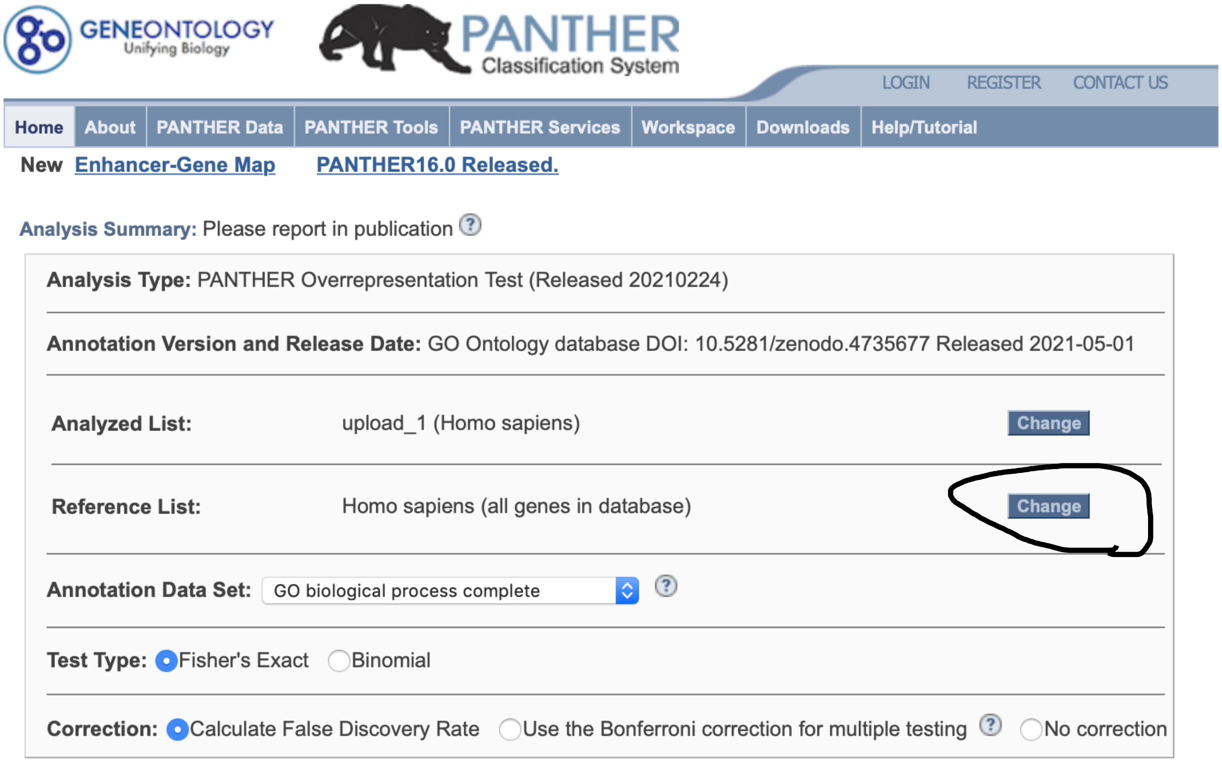
Khdrbs2
Rnf149
Wdr75
Pgap1
Hspd1
Mob4

biological process

Homo sa

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs

“Launch”



New Enhancer-Gene Map PANTHER16.0 Released.

Analysis Summary: Please report in publication ?

Analysis Type: PANTHER Overrepresentation Test (Released 20210224)

Annotation Version and Release Date: GO Ontology database DOI: 10.5281/zenodo.4735677 Released 2021-05-01

Analyzed List: upload_1 (Homo sapiens) Change

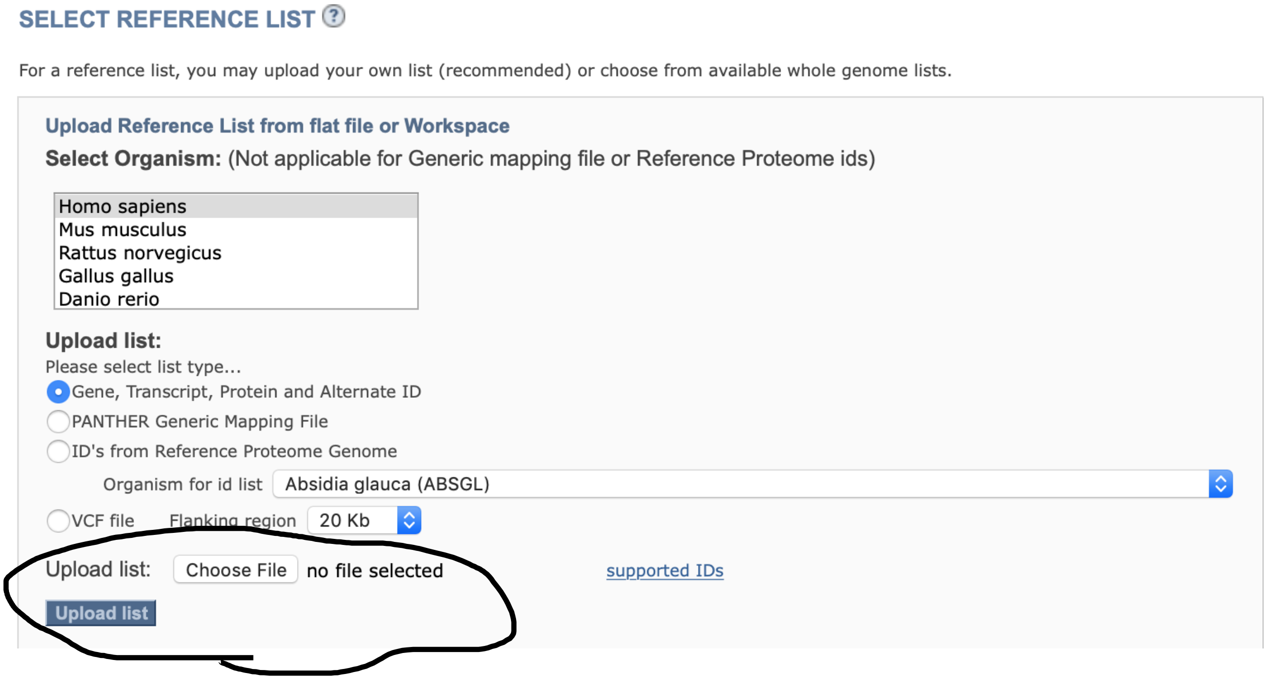
Reference List: Homo sapiens (all genes in database) Change

Annotation Data Set: GO biological process complete ?

Test Type: ● Fisher's Exact ○ Binomial

Correction: ● Calculate False Discovery Rate ○ Use the Bonferroni correction for multiple testing ? ○ No correction

- Make sure background gene set is in a text file



SELECT REFERENCE LIST ?

For a reference list, you may upload your own list (recommended) or choose from available whole genome lists.

Upload Reference List from flat file or Workspace

Select Organism: (Not applicable for Generic mapping file or Reference Proteome ids)

Homo sapiens
Mus musculus
Rattus norvegicus
Gallus gallus
Danio rerio

Upload list:

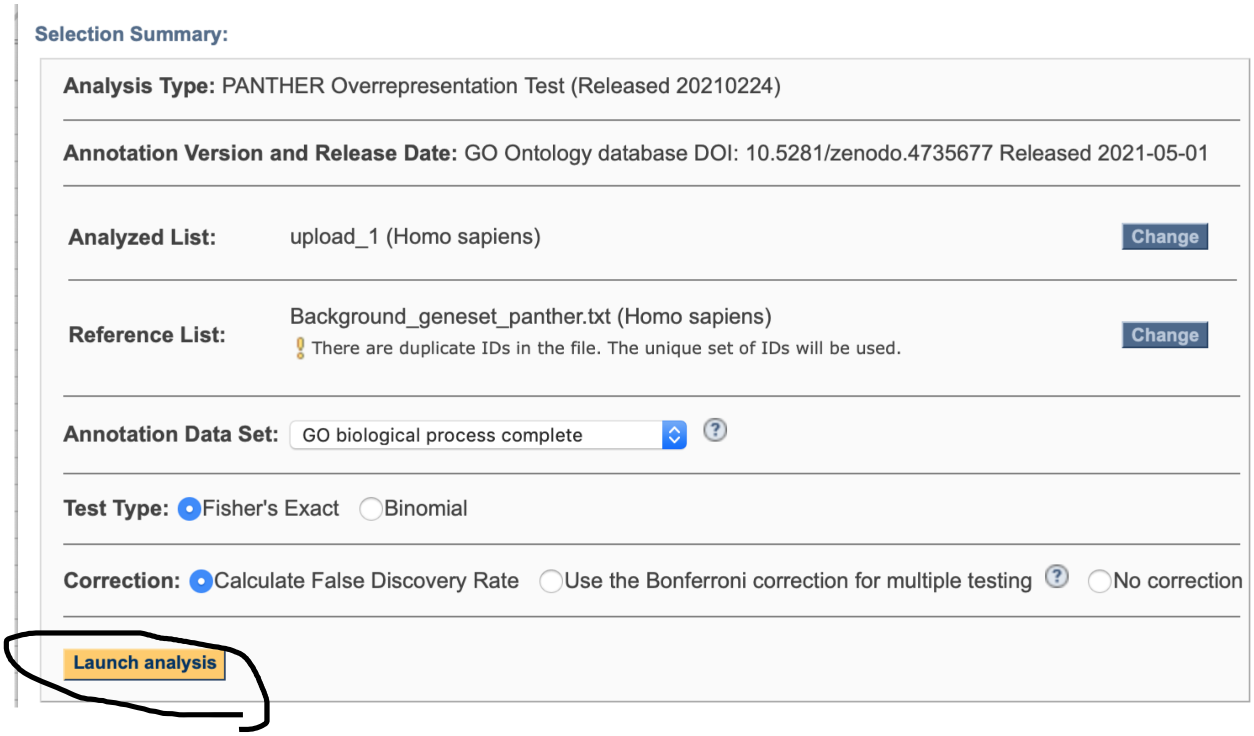
Please select list type...

● Gene, Transcript, Protein and Alternate ID
○ PANTHER Generic Mapping File
○ ID's from Reference Proteome Genome

Organism for id list: Absidia glauca (ABSGL)

○ VCF file Flanking region 20 Kb

Upload list: Choose File no file selected supported IDs



Selection Summary:

Analysis Type: PANTHER Overrepresentation Test (Released 20210224)

Annotation Version and Release Date: GO Ontology database DOI: 10.5281/zenodo.4735677 Released 2021-05-01

Analyzed List: upload_1 (Homo sapiens) Change

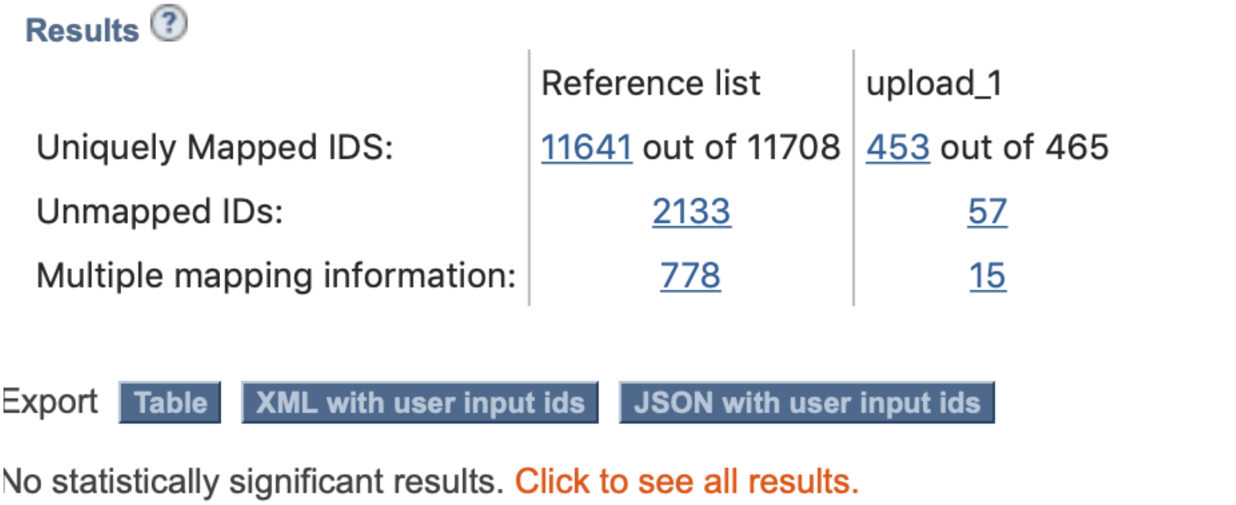
Reference List: Background_geneset_panther.txt (Homo sapiens) Change
! There are duplicate IDs in the file. The unique set of IDs will be used.

Annotation Data Set: GO biological process complete ?

Test Type: ● Fisher's Exact ○ Binomial

Correction: ● Calculate False Discovery Rate ○ Use the Bonferroni correction for multiple testing ? ○ No correction

Launch analysis



Results ?

	Reference list	upload_1
Uniquely Mapped IDS:	11641 out of 11708	453 out of 465
Unmapped IDs:	2133	57
Multiple mapping information:	778	15

Export Table XML with user input ids JSON with user input ids

No statistically significant results. Click to see all results.