Lab07

Gene Enrichment Analysis.

Also called Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a pre-defined set of genes (eg: those belonging to a specific GO term or KEGG pathway) shows statistically significant, concordant differences between a sample and a reference (two biological states).

A few websites:

- https://geneontology.org/
- https://www.ebi.ac.uk/QuickGO/
- http://bioinformatics.sdstate.edu/go/
- https://useast.ensembl.org/index.html
- https://www.pantherdb.org/

Document all work during the dry lab for each exercise, defining all your tools and input parameters, data output, and interpretation.

From Lab07, there are two Assignments to be submitted to myCourses:
Discussion 7.1
Activity 7.1

ShinyGO: tool for enrichment analysis using gene lists.
http://bioinformatics.sdstate.edu/go/

ShinyGO (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7178415/) is based on a large annotation database derived from Ensembl and STRING-db for 59 plant, 256 animal, 115 archeal and 1678 bacterial species. ShinyGO is a Shiny application (https://github.com/rstudio/shiny) developed based on several R/Bioconductor packages, and a large annotation and pathway database compiled from many sources.

As an example, we will analyze a set of 149 genes (Lab07-lymphoblasts_GeneIDs) up-regulated in lymphoblasts cells (TK6, WTK1, and NH32) treated with ionizing radiation: The list is also uploaded into myCourses>Content>Labs.

ADORA3 AHR AKAP5 ARL2BP ATF5 AXL BAI3 BATF BBC3 BIRC3 BTG1 BTG2 CARTPT CASP1 CCL18 CCL20 CCL21 CCL3 CCL4 CCNE2 CCNF CCNG1 CCNG2 CCNI CCNK CCR1 CCT4 CCT7 CD164 CD59 CD79A CD80 CD83 CDK1 CDK6 CDKN1A CDKN1C CFLAR CGRRF1 CLK1 COL4A2 COL6A1 COX7B CRAT CXCL10 DDB2 DDIT3 DKC1 DLX2 DPYSL3

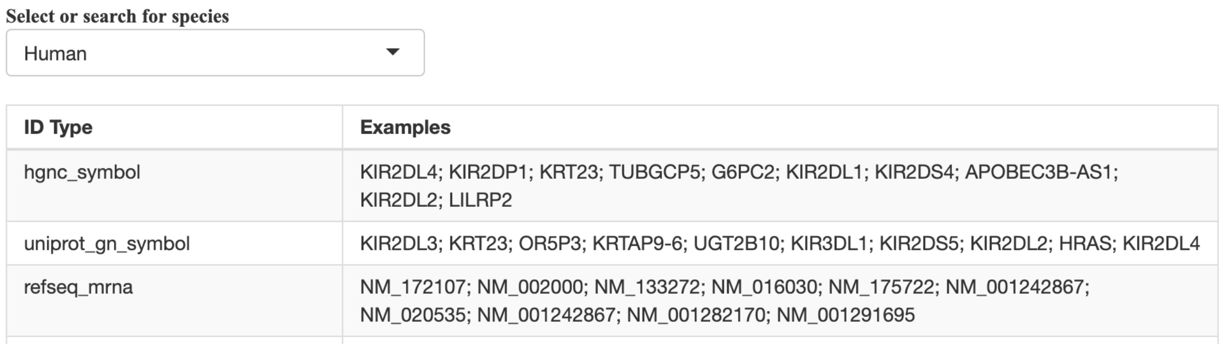
DUSP6 ENC1 ENO1 EP300 FAM134C FAS FCN3 FGFR1 FN1 GDF15 GRN HBEGF HOXC5 IFITM1 IFNGR1 IL2RB IRF4 IRF9 ITGA2 JAK2 JUNB KIF20B KLF4 KLRC2 KRT14 KRT19 KRT8 LGALS3 LZTR1 MAP7 MAX MDFI MED21 MSC MSH2 MYC NCOA3 NDUFB5 NFE2L1 NFKB1 NFKB2 NFKBIA NFKBIE NR4A3 PCNA PKD2L1 PLAGL2 PLAU PLCG2 PLK1 PLOD3 POU3F4 POU5F1 PPM1D PRMT1 PTN PTPN11 PTPN22 PTPN6 PTTG1 RABGAP1 RAD50 RANBP1 RB1 RELA RFX5 ROCK1 RORA RPS6KA1 RRM2B SGK1 SIPA1 SMAD3 SNAPC2 STAT1 STAT5A STK4 TANK TCF21 TFDP2 TGFBR3 TGIF1 TLE3 TNFAIP3 TNFAIP8 TNFRSF8 TNFSF10 TP53I3 TP53TG1 TRAF1 TRRAP TSC22D3 UBE2C VEGFC XPC ZBTB48 ZNF141 ZNF274 ZNF85

Before you do anything, let's take a look at the website and all the elements in there: http://bioinformatics.sdstate.edu/go/

Hover over with your cursor to see descriptions of:

- Change species
- Demo genes
- Background
- Pathway database (shows up after you paste your genes and click Submit)
- FDR cutoff
- Pathways

Now, click "Gene IDs examples" and select Human (you will have to wait a few seconds for it to load). What do represent the different ID symbols? What does it imply regarding having a GO ontology database (knowledgebase) with genes associated with GO terms?

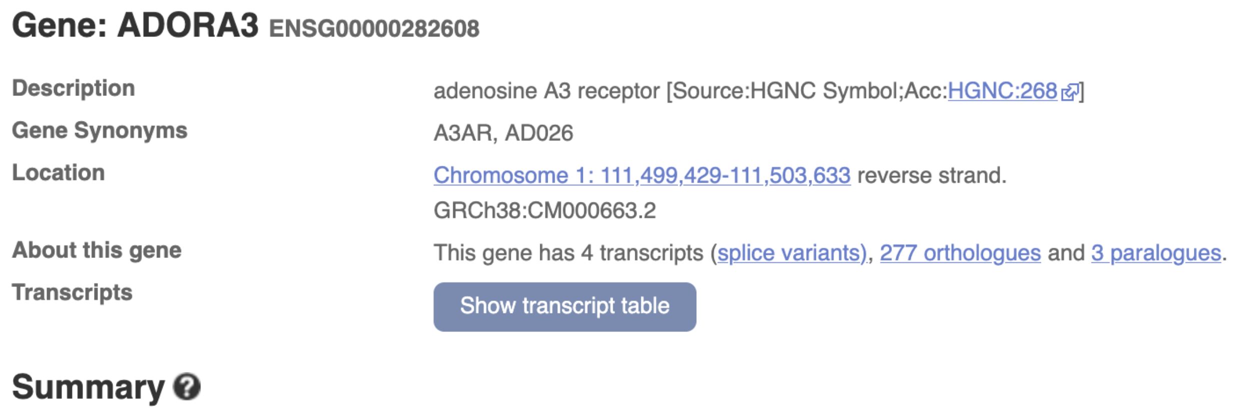


Select or search for species

Human

ID Type	Examples
hgnc_symbol	KIR2DL4; KIR2DP1; KRT23; TUBGCP5; G6PC2; KIR2DL1; KIR2DS4; APOBEC3B-AS1; KIR2DL2; LILRP2
uniprot_gn_symbol	KIR2DL3; KRT23; OR5P3; KRTAP9-6; UGT2B10; KIR3DL1; KIR2DS5; KIR2DL2; HRAS; KIR2DL4
refseq_mrna	NM_172107; NM_002000; NM_133272; NM_016030; NM_175722; NM_001242867; NM_020535; NM_001242867; NM_001282170; NM_001291695

Before engaging in the Enrichment Analysis, open Ensembl, a genome browser for vertebrate genomes, in a new window https://useast.ensembl.org/index.html and check the website. Search for one of the genes in our list (eg. ADORA3 in Human), and check all the names and GeneIDs associated with the gene entry:



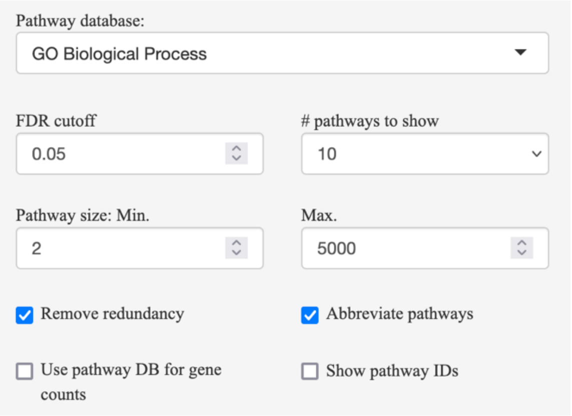
Gene: ADORA3 ENSG00000282608

Description	adenosine A3 receptor [Source:HGNC Symbol;Acc:HGNC:268]
Gene Synonyms	A3AR, AD026
Location	Chromosome 1: 111,499,429-111,503,633 reverse strand. GRCh38:CM000663.2
About this gene	This gene has 4 transcripts (splice variants), 277 orthologues and 3 paralogues.
Transcripts	Show transcript table

Summary

Name ADORA3 (HGNC Symbol)
MANE
UniProtKB identifiers: P0DMS8
RefSeq
CCDS
LRG
Ensembl version ENSG00000282608.2

Back to ShinyGO… Enrichment Analysis. Paste the lymphoblasts gene list and click Submit. Then, it will appear the "Pathway database" to select from. Select **GO Biological Process**.



Pathway database: GO Biological Process

FDR cutoff: 0.05

pathways to show: 10

Pathway size: Min. 2

Max. 5000

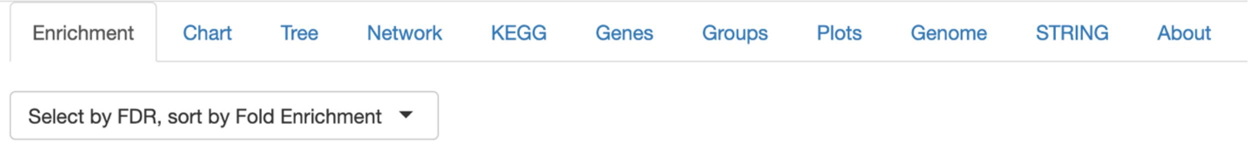
☑ Remove redundancy

☑ Abbreviate pathways

☐ Use pathway DB for gene counts

☐ Show pathway IDs

You can check your results on the right-side panel and modify the settings (FDR, #pathways to show…) on the go. Browse all the options for visualization:



For the assignment, you will need to combine a few screenshots, using ten #pathways to show, with the results of:

- Enrichment table
- Chart
- Network

Feel free to browse other GO roots, such as molecular function, cellular components, and even KEGG ontology database.
As you can see, you can modify graphs based on **FDR and Fold Enrichment**. FDR is adjusted from the hypergeometric test. **Fold Enrichment (FE)** is defined as the proportion of genes in your list belonging to a pathway, divided by the corresponding proportion in the background. FDR tells us how likely the enrichment is by chance; **Fold Enrichment** indicates how drastically genes of a certain pathway is overrepresented. While FDR measures statistical significance, fold enrichment indicates effect size.

$$FE = \frac{A}{B}$$

$$A = {}^{Pathway\ genes\ set}/_{Set\ genes}$$

$$B = {}^{Pathway\ genes}/_{Background\ genes}$$

Using the fold enrichment values in the table, can you calculate how many (human) genes are used here as background?

Discussion 7.1
Please write down the steps and parameters (methods) you used, describing all the components, tables/charts, and what they represent. (Once you finish, submit the report with explanations to myCourses (in Assignments).

Gene sets access

Where can I get a specific gene set enriched for GOs? You might ask. Well, almost every paper with some transcriptome analysis between samples with different treatments, time points, and environmental conditions is a good starting point. But remember you need a reference (background) with GO annotations mapped to corresponding Gene IDs.

A good source for gene sets enrichment analysis is MSigDB (https://www.gsea-msigdb.org), where you can browse different human (and mouse) gene sets associated with signatures. For instance, there are **Hallmark gene sets**, which are associated with some specific activity.

You will test one of these hallmark gene sets using the same previous gene enrichment analysis steps in ShinyGO. Find your preferred hallmark collection in **https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=H** and save the gene set. On the description page, look over the details. Under **Download gene set**, select “grp” and save it into your computer. You can open the file with a text editor. You will see something like

```
HALLMARK_DNA_REPAIR
# https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/HALLMARK_DNA_REPAIR
AAAS
ADA
ADCY6
ADRM1
…
```

Would you like to test another specific gene set from another species you are working on? If ShinyGO has the species annotations in its collection (under Change species), feel free to do so!

Copy/Paste the gene set you choose into ShinyGO and proceed to visualize the enriched pathways (GO BP, MF and CC).

Activity 7.1

Once you finish, submit the report to myCourses (in Assignments). Please write down the steps and parameters (methods) you used, describing all the components, tables/charts, and what they represent. You only need to “screenshot” one pathway database for the activity.