

Over-representation analysis (ORA) practical lab (Module2)

Veronique Voisin

You can choose to do these exercises using the questions as your only guide - or see the following pages for the step-by-step checklist to finding these answers.

The data set used for this practical lab contains transcriptomics data obtained from MCF7 cells, a human breast cancer line, treated or non treated with estradiol. The cells were treated with estradiol for 12, 24 or 48 hours. Total RNA extracted from the cells was amplified, labeled and hybridized to Affymetrix GeneChip U133 Plus 2.0 microarrays. The data are available in the Gene Expression Omnibus (GEO) repository under the accession number GSE11352 (PMID: [17542648](https://pubmed.ncbi.nlm.nih.gov/17542648/)). The practical lab contains two exercises. Exercise 1 uses GSEA (<http://www.broadinstitute.org/gsea/index.jsp>) to perform gene-set enrichment analysis and exercise 2 uses g:Profiler (<http://biit.cs.ut.ee/gprofiler/>).

Exercise 1:

For this exercise, our goal is to upload the 3 required files into GSEA, set up the parameters, run GSEA, open and explore the gene-set enrichment results. We use as input file for GSEA the normalized data for all samples included in the GSE11352 dataset and formatted as a '.gct' file. GSEA will assess the amplitude of differential gene expression levels between the two groups of interest, in this case the treated samples and non treated samples at 12 hours using a t-test for each gene. The '.cls' file tells GSEA which samples correspond to our groups of interest. GSEA ranks the genes based on t values from the t-test and performs the gene-set enrichment analysis using a modified Kolmogorov-Smirnov statistics. The output result folder contains several files, and two of them are the summary tables displaying enrichment statistics for each gene-set (pathway) that has been tested and contained in the provided '.gmt' file. The '.gmt' file (gene-set file) provided for this exercise contains gene-sets obtained from KEGG, MsigDB-c2, NCI, Biocarta, IOB, Netpath, HumanCyc, Reactome and the Gene Ontology (GO) databases. (<http://baderlab.org/GeneSets>).

Before starting this exercise, launch GSEA using the instructions provided on the wiki and download the 3 required files:

- [MCF7_Expression_matrix.gct](#)
- [MCF7_groups.cls](#)
- [Human_GO_AllPathways.gmt](#)

Exercise 2:

For this exercise, we are going to use a list of 428 genes that are differentially expressed in the MCF7 cells treated with estradiol for 24hr compared to the control samples. Our goal is to perform gene-set enrichment on this list using the g:Profiler tool and to explore the results. The Gene Ontology Biological Process, the KEGG and Reactome are going to be used as the pathway databases. g:Profiler uses a Fisher's exact test to calculate the significance of the gene-set enrichment.

Before starting this exercise, download the required file from the CBW wiki :

- [24hr_topgenes.txt](#) .

EXERCISE 1:

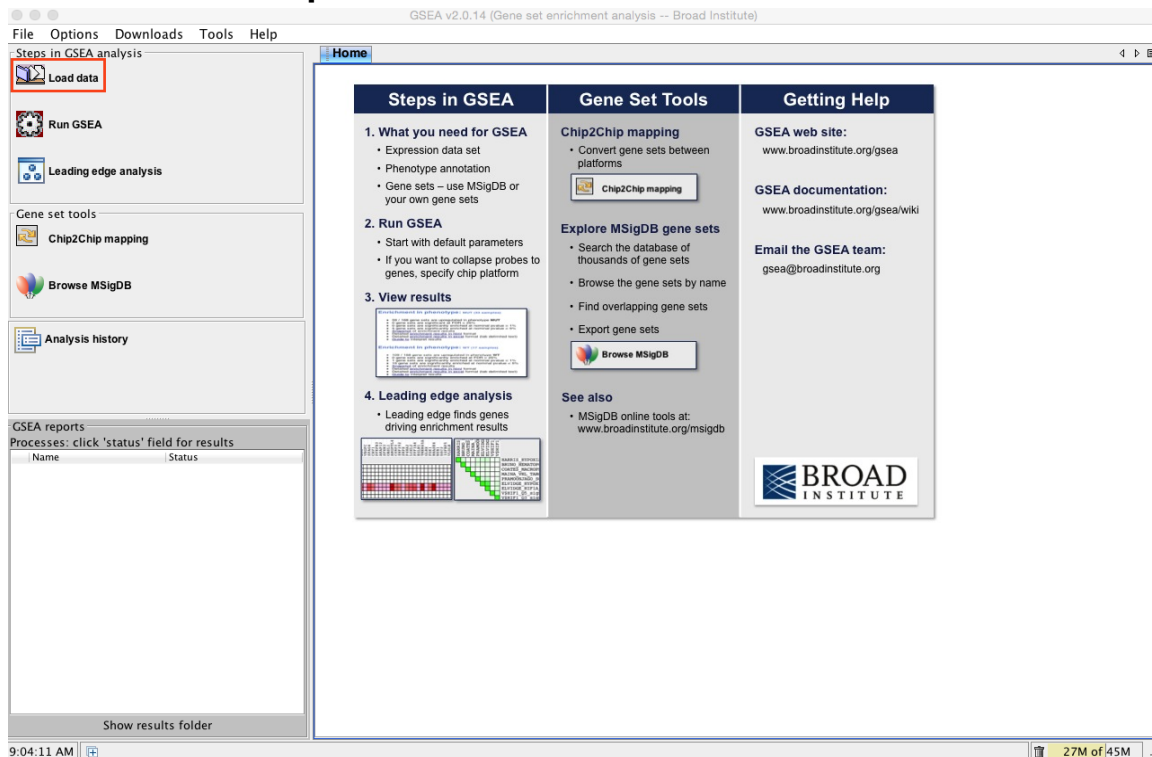
Step	Action	Check
1	Launch GSEA.	
2	Locate the ' Load data ' icon at the upper left corner of the window and click on it .	
3	<p>In the central panel, select 'Method 1' and 'Browse for files'. A new window pops up. Browse your computer to locate the 3 files : Import the MCF7_Expression_matrix.gct, MCF7_groups.cls and Human_GO_AllPathways.gmt. Click on 'Choose'. A message pops up when the files are loaded successfully. Click on 'OK'.</p> <p>Alternatively, you can choose 'Method 3' to 'drag and drop files here'; you need to click on the 'Load these files!' button in this case.</p>	
4	Locate the ' Run GSEA ' icon at the left side of the main window located below the ' Load data button ' and click on it .	
5	In the central window called ' GSEA: Set parameters and run enrichment tests ', fill the first field called ' Expression dataset ' by clicking on the up and down arrows. Choose MCF7_Expression_matrix . Tip: Mousing over the parameters fields will highlight a short description.	
6	Click on the 3 dots [...] of the radio button corresponding to the ' Gene sets database ' field. A new window will pop up after approximately 10 seconds. Using the right arrow in the menu bar of this window, locate the ' Gene Matrix (local gmx/gmt) ' tab and select the file Human_GO_AllPathways.gmt . Click on ' OK '.	
7	<p>In the field 'Number of permutations' enter the number 100.</p> <p>Note: for this exercise and purpose of demonstration, please use 100. For real life data analysis, 2000 permutations is recommended and it will require about 1 hour to run using a complete set of gene-sets.</p>	
8	In the ' Phenotype labels ' field, click on the 3 dots [...] of the radio button. A window " Select a phenotype " will pop up. Make sure that the file MCF7_groups.cls appears as selected source file; locate and select the comparison ES12_versus_NT12 . Click on ' OK '.	
9	Set the ' Collapse dataset to gene symbols ' field to ' false '.	
10	Set ' Permutation ' type to ' gene-set '.	
11	Leave the ' Chip platform(s) ' empty.	
12	In ' Basic Fields ', choose an informative name for your analysis in the ' Analysis name field '. Tip: name of the comparison that you are making and date (e.g ES12_versus_NT12_date).	
13	Set the ' Metric for ranking genes ' to ' tTest '.	

14	'In the Save results in this folder' , click on the 3 dots [...] and browse your computer to select a folder.	
15	Locate and click the 'Run' button at the bottom right corner of the window. Tip: you may need to expand the window to see the Run button. Note: it takes 5 min to run using a maximum of 1.4Gb of memory. GSEA has finished to run when a message 'Success 5' appears in the Status field of the GSEA reports box.	
16	In the 'GSEA reports' box, click on 'Success 5' to see the results. Open the links and explore the results.	

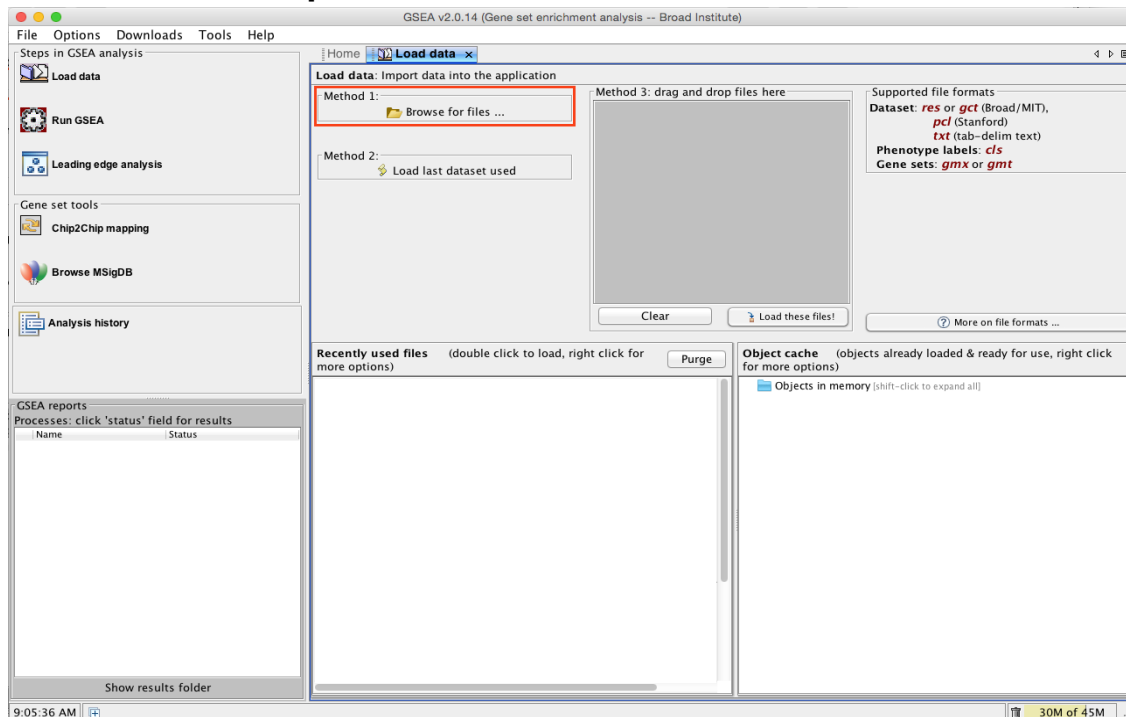
EXERCISE 2

Step	Action	Check
1	Go to g:Profiler 's homepage at http://biit.cs.ut.ee/gprofiler/	
2	Ensure that 'Organism' is set to 'Homo sapiens'	
3	In the 'Query' box, copy and paste the 428 genes listed in the file 24hr_topgenes.txt	
4	Select the following 'Options' by checking the corresponding boxes: <ul style="list-style-type: none"> • 'Significant only' • 'No electronic GO annotations' • 'Hierarchical sorting' 	
5	Select the following gene-set databases by checking the corresponding boxes: <ul style="list-style-type: none"> • 'Gene Ontology' 'Biological Process' • 'Biological pathways' 'KEGG' and 'Reactome' 	
6	Click on 'Show advanced options' .	
7	Select the following 'advanced options': <ul style="list-style-type: none"> • 'Size of functional category' : 3 (min) and 500 (max) • 'Size of Q&T' : min of 2 	
8	Click on the 'g:Profile!' button.	
9	Click on the warning message 'Some gene identifiers are ambiguous. resolve these manually?' Select the first <i>ENSEMBL ID</i> for each gene and click on 'Resubmit query' .	
10	Explore the results. Which term has the best corrected p-value? Which genes in our list are included in this term?	
11	If time permits, play with input parameters, e.g. add 'TRANSFAC TFBS' and 'miRBase microRNAs' databases, rerun the query by clicking on the 'g:Profile!' button and explore the new results.	

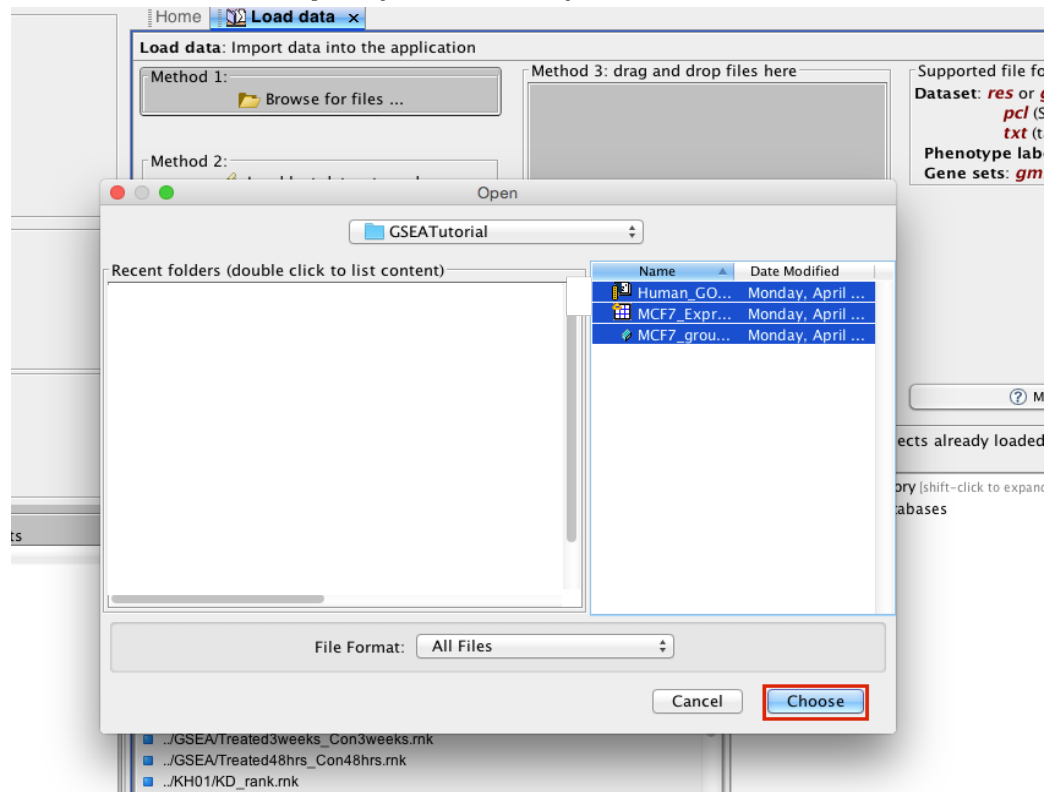
EXERCISE 1: Steps 1-2



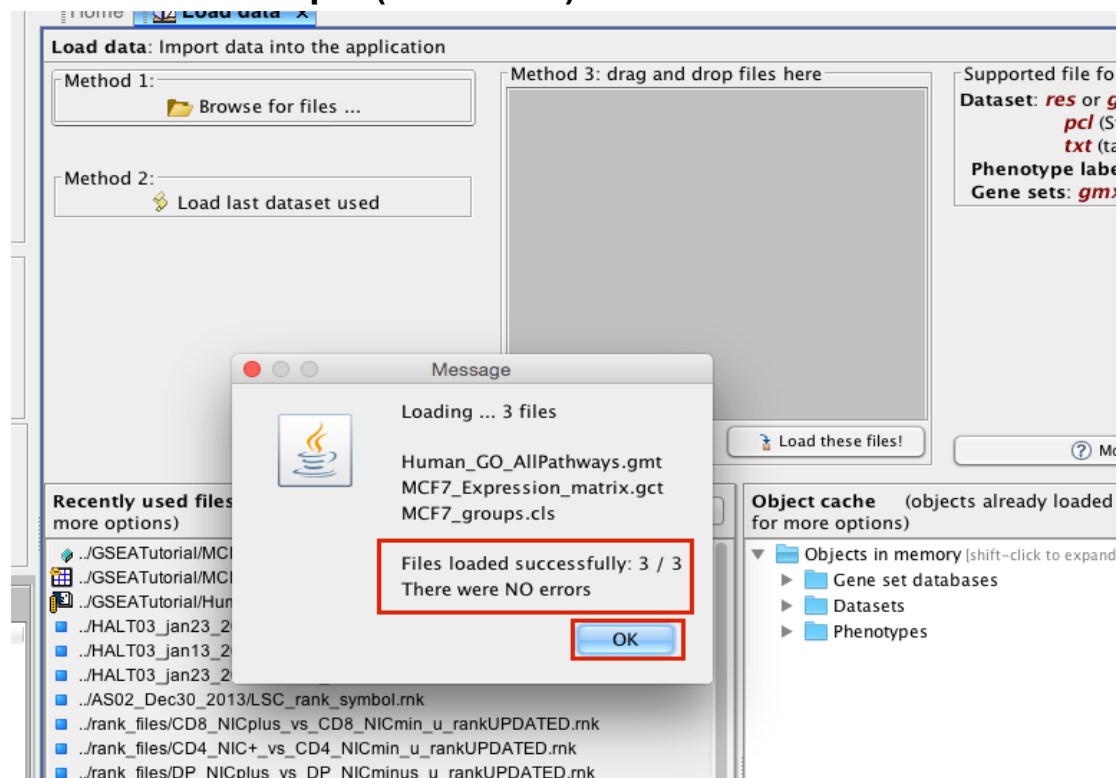
EXERCISE 1: Step 3



EXERCISE 1: Step 3 (continued)



EXERCISE 1: Step 3 (continued)



EXERCISE 1: Step 4

GSEA v2.0.14 (Gene set enrichment analysis -- Broad Institute)

File Options Downloads Tools Help

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis

Gene set tools

- ChIP2ChIP mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Load data: Import data into the application

Method 1: Browse for files ...

Method 2: Load last dataset used

Method 3: drag and drop files here

Supported file formats

Dataset: *res* or *gct* (Broad/MIT), *pcl* (Stanford), *txt* (tab-delim text)

Phenotype labels: *cls*

Gene sets: *gmx* or *gmt*

Recently used files (double click to load, right click for more options)

- .GSEATutorial/MCF7_groups.cls
- .GSEATutorial/MCF7_Expression_matrix.gct
- .GSEATutorial/Human_GO_AllPathways.gmt
- .HALT03_jan23_2014/bayseqNR_uRANK.mk
- .HALT03_jan23_2014/bayseqR_jan17_u_RANK.mk
- .HALT03_jan23_2014/treated_uRANK.mk
- .JAS02_Dec30_2013/LSC_rank_symbol.mk
- .rank_files/CD8_NICplus_vs_CD8_NICmin_u_rankUPDATED.mk
- .rank_files/CD4_NIC+_vs_CD4_NICmin_u_rankUPDATED.mk
- .rank_files/OP_NICplus_vs_DP_NICminus_u_rankUPDATED.mk
- .GSEA/Treated9days_Con9days.mk
- .GSEA/Treated48hrs_Con0hr.mk
- .GSEA/Treated3weeks_Con3weeks.mk
- .GSEA/Treated48hrs_Con48hrs.mk
- .KH01/KD_rank.mk
- .KH01/OE_rank.mk
- .JG04_RBPJ_bindingsites_Dec2_2014/CD4_NIC+_vs_CD4_NICmin_u_rank
- .JG07_part3_proteomics/proteomics_rank.mk
- .JHR_targets_april27/OE_rank.mk
- .JHR_targets_april27/KD_rank.mk

Purge

Object cache (objects already loaded & ready for use, right click for more options)

- Objects in memory (shift-click to expand all)
 - Gene set databases
 - Datasets
 - Phenotypes

9:28:50 AM 4730 [INFO] Loading ... 3 files: Human_GO_AllPathways.gmt MCF7_Expression_matrix.gct MCF7_groups.cls Files loaded successfully: 3 / 3 There were NO err... 122M of 217M

EXERCISE 1: Step 5

GSEA v2.0.14 (Gene set enrichment analysis -- Broad Institute)

File Options Downloads Tools Help

Steps in GSEA analysis

- Load data
- Run GSEA**
- Leading edge analysis

Gene set tools

- ChIP2ChIP mapping
- Browse MSigDB

Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

GSEA: Set parameters and run enrichment tests

Required fields

Expression dataset: MCF7_Expression_matrix [20326x18 (ann: 20326,18,chip na)]

Gene sets database: ...

Number of permutations: 1000

Phenotype labels: ...

Collapse dataset to gene symbols: true

Permutation type: phenotype

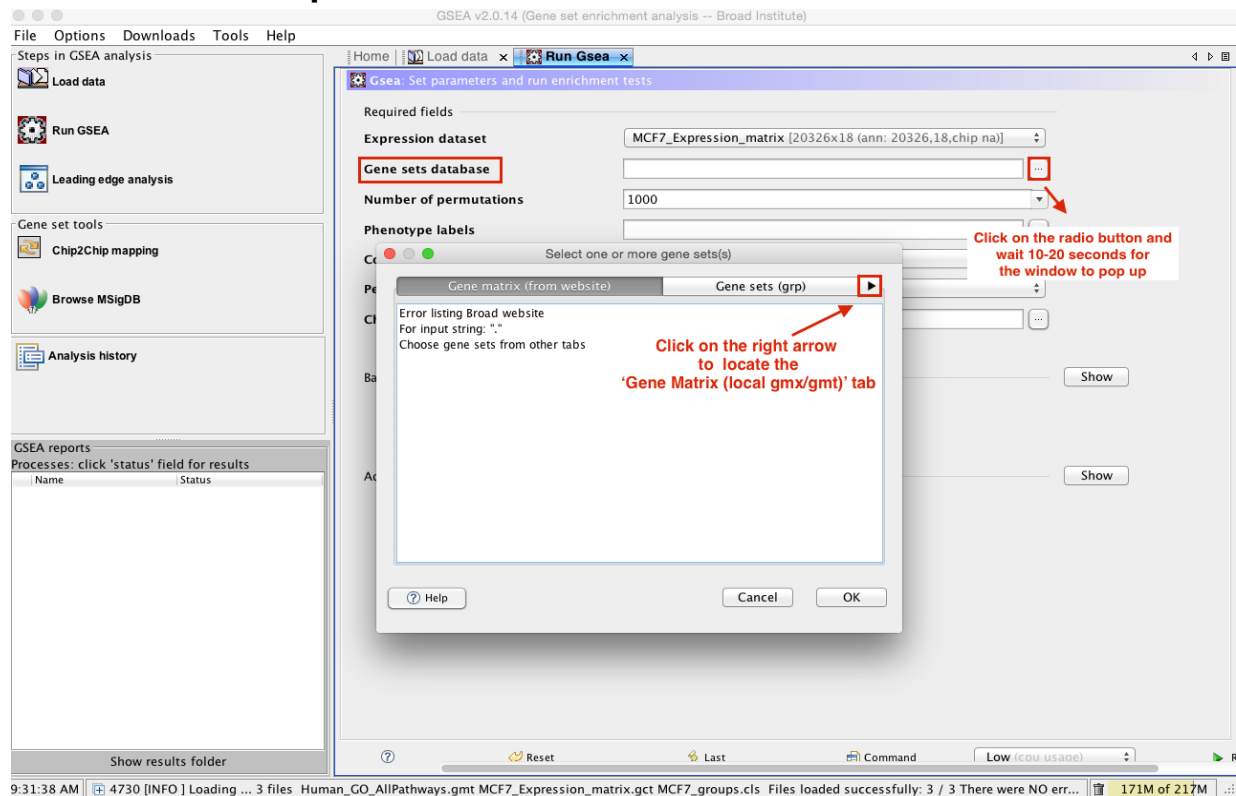
Chip platform(s): ...

Basic fields: Show

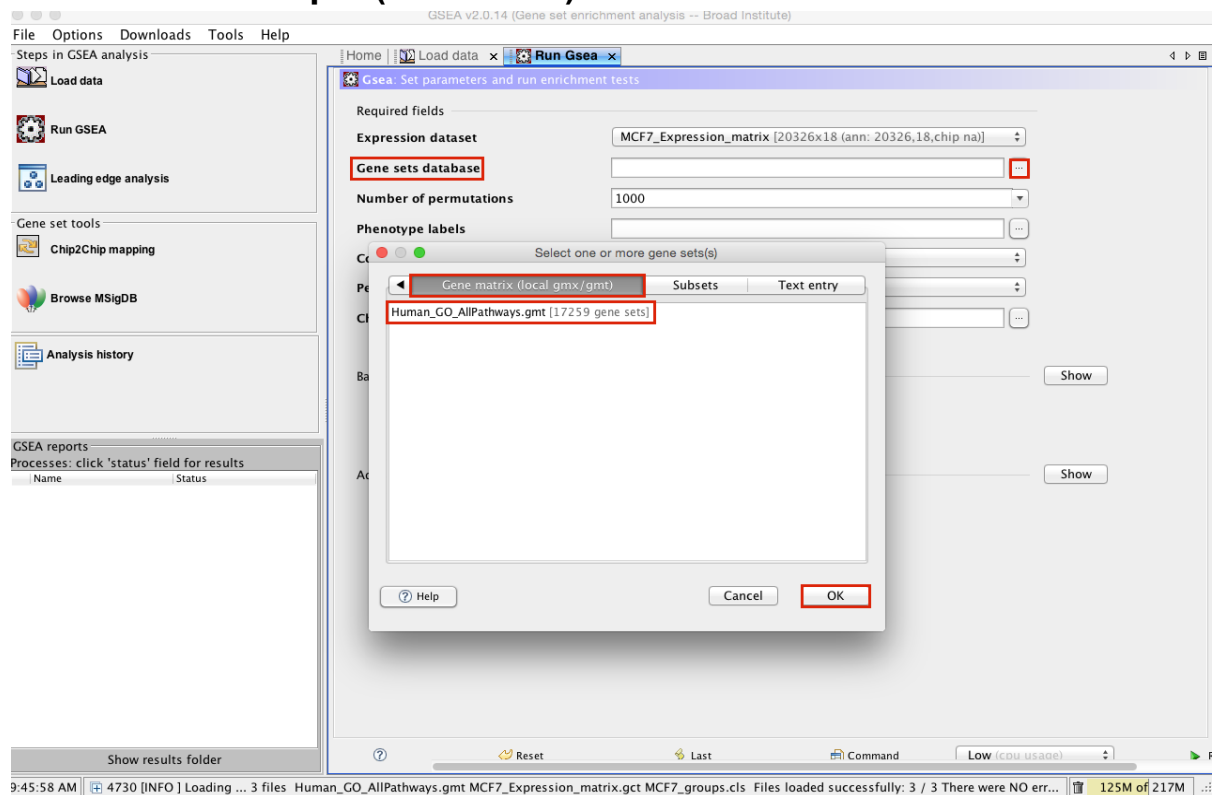
Advanced fields: Show

Reset Last Command Low (cpu usage)

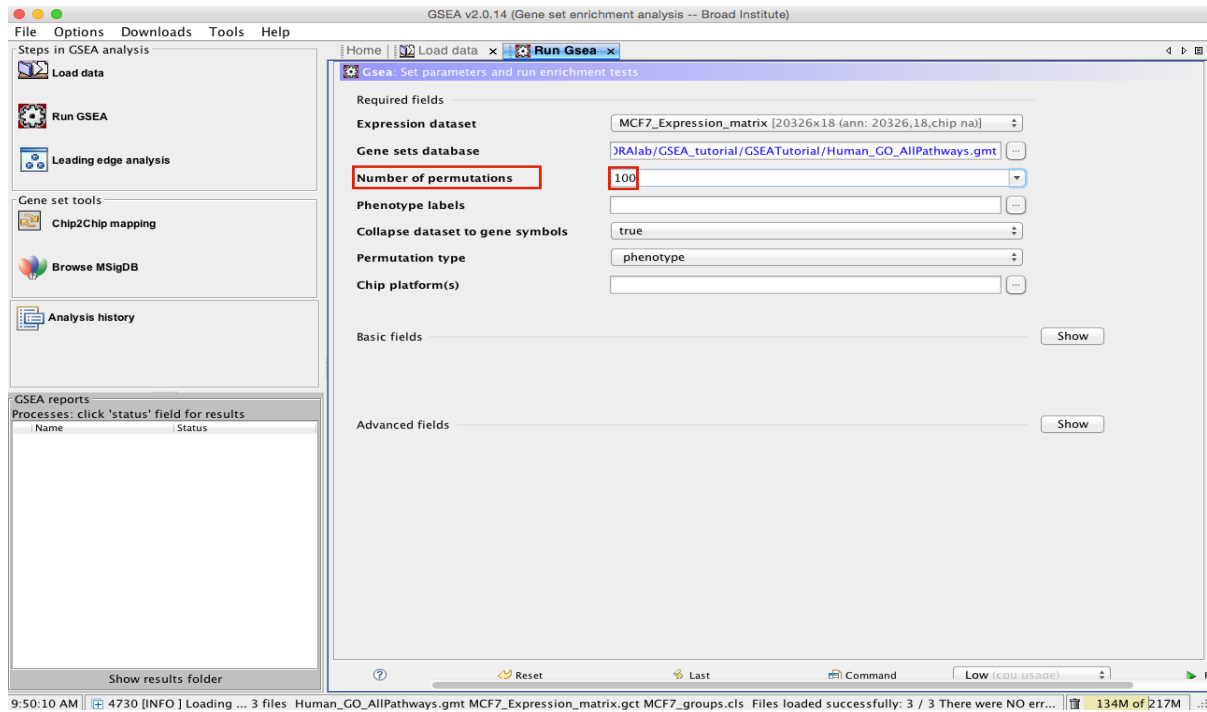
EXERCISE 1: Step 6



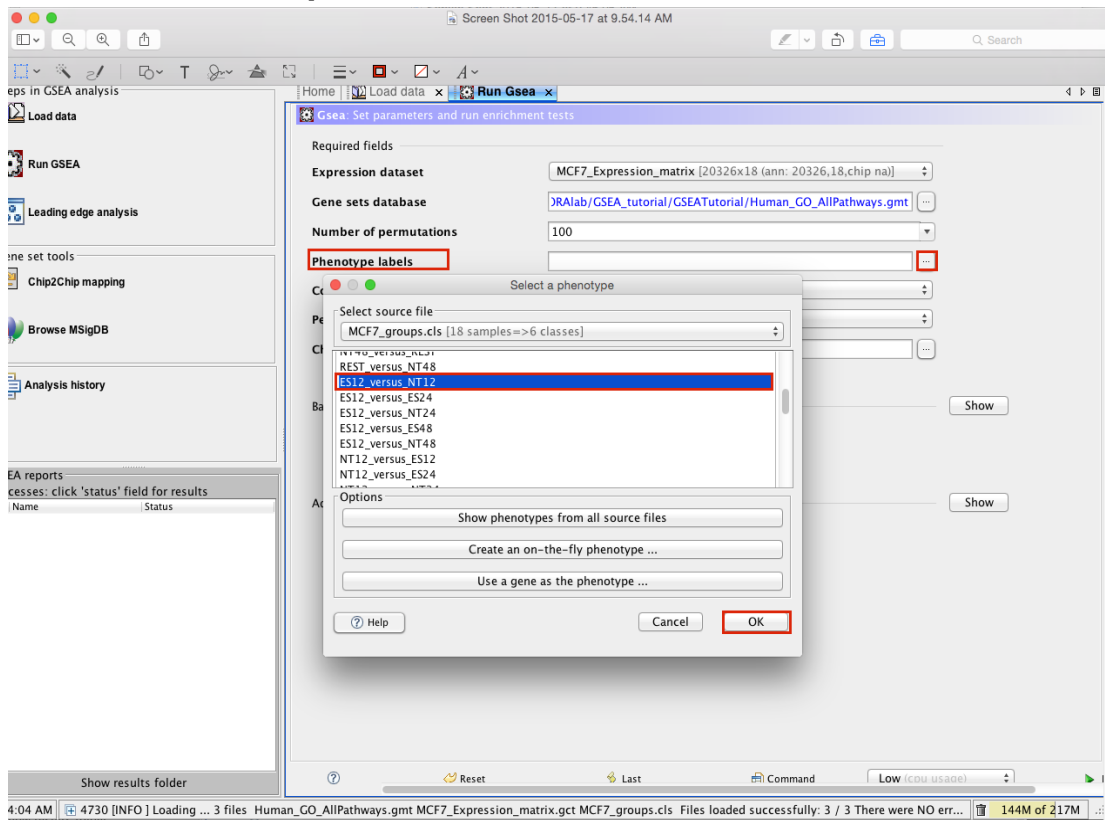
EXERCISE 1: Step 6 (continued)



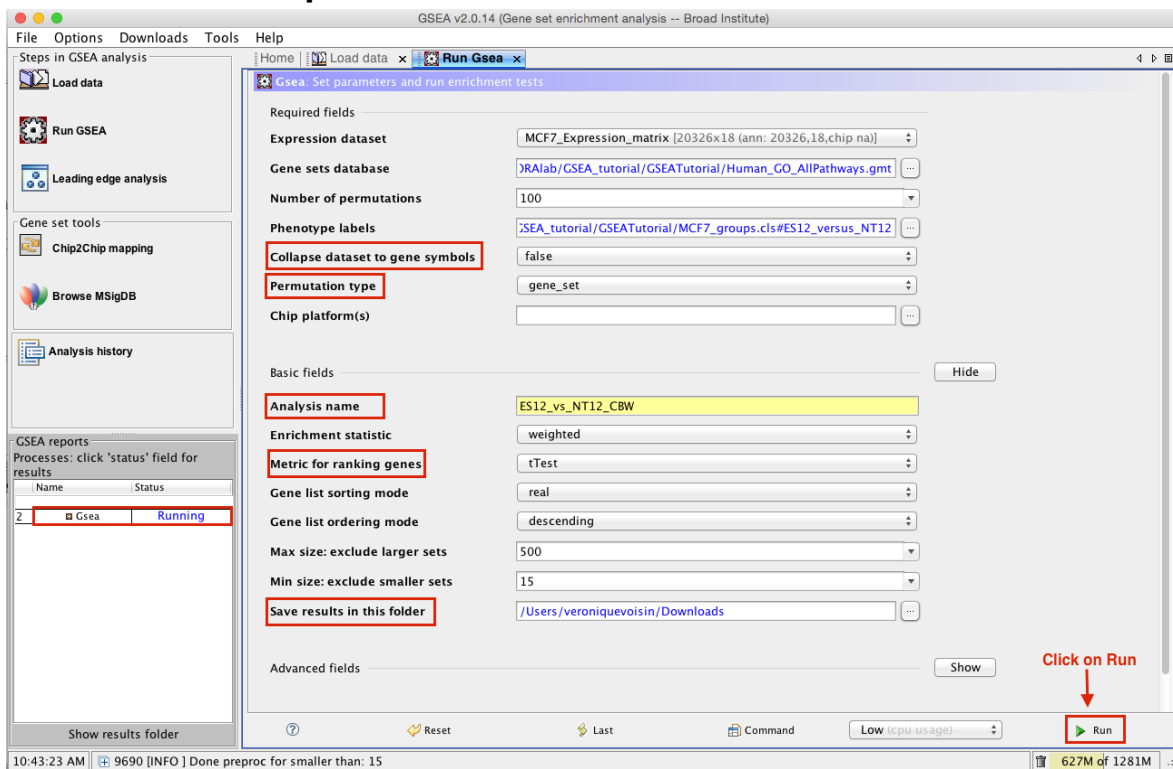
EXERCISE 1: Step 7



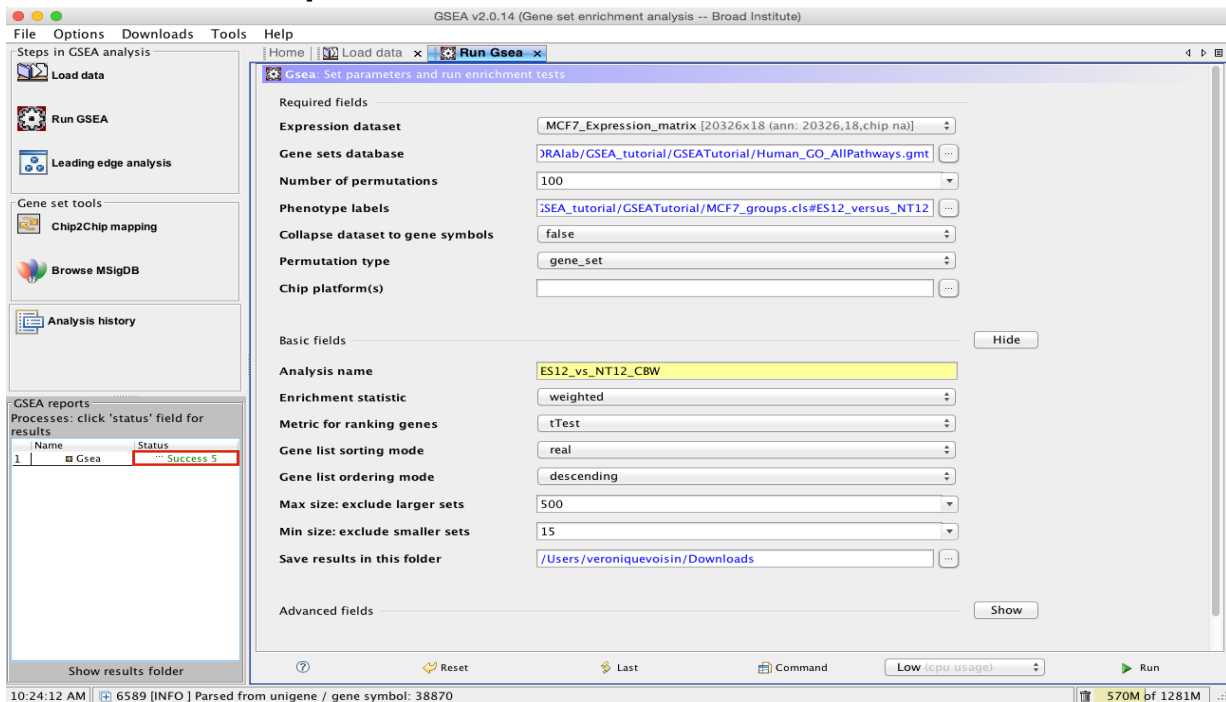
EXERCISE 1: Step 8



EXERCISE 1: Steps 9-15



EXERCISE 1: Step 16



EXERCISE 1: Step 17

GSEA Report for Dataset MCF7_Expression_matrix

Enrichment in phenotype: ES12 (3 samples)

gene-sets enriched in genes up-regulated in treated cells compared to non-treated samples

- 2120 / 4756 gene sets are upregulated in phenotype **ES12**
- 665 gene sets are significant at FDR < 25%
- 422 gene sets are significantly enriched at nominal pvalue < 1%
- 612 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

gene-sets enriched in genes down-regulated in treated cells compared to non-treated samples

Enrichment in phenotype: NT12 (3 samples)

- 2636 / 4756 gene sets are upregulated in phenotype **NT12**
- 445 gene sets are significant at FDR < 25%
- 337 gene sets are significantly enriched at nominal pvalue < 1%
- 601 gene sets are significantly enriched at nominal pvalue < 5%
- [Snapshot](#) of enrichment results
- Detailed [enrichment results in html](#) format
- Detailed [enrichment results in excel](#) format (tab delimited text)
- [Guide to](#) interpret results

Dataset details


- The dataset has 20323 features (genes)
- No probe set => gene symbol collapsing was requested, so all 20323 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 12503 / 17259 gene sets
- The remaining 4756 gene sets were used in the analysis
- List of [gene sets used and their sizes](#) (restricted to features in the specified dataset)

Gene markers for the ES12 *versus* NT12 comparison

- The dataset has 20323 features (genes)
- # of markers for phenotype **ES12**: 9758 (48.0%) with correlation area 49.7%
- # of markers for phenotype **NT12**: 10565 (52.0%) with correlation area 50.3%
- Detailed [rank ordered gene list](#) for all features in the dataset
- [Heat map and gene list correlation](#) profile for all features in the dataset



[g:GOST](#) Gene Group Functional Profiling
[g:Cocoa](#) Compact Compare of Annotations
[g:Convert](#) Gene ID Converter
[g:Sorter](#) Expression Similarity Search
[g:Orth](#) Orthology search

[Welcome!](#)
[About](#)
[Contact](#)
[Beta](#)
[Archives](#)
[R](#)

J. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) NAR 35 W193-W200 [PDF]

J. Reimand, T. Arak, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2011 update) Nucleic Acids Research 2011; doi: 10.1093/nar/gkr378 [PDF]

[?] Organism

Homo sapiens

[?] Query (genes, proteins, probes, term)

BOP1
F11R
CNP
RITMK
ASS1
IRX2
LOC100133166
SLC2A1
THBS1

[?] or Term ID:

g:Profile! Clear

Example or random query

Options

☒ Significant only

☐ Ordered query

☐ No electronic GO annotations

☐ Chromosomal regions

☒ Hierarchical sorting

☐ Hierarchical filtering

Show all terms (no filtering)

☐ Output type

Graphical (PNG)

Hide advanced options

☐ Measure underrepresentation

☐ Gene list as a stat. background

☐ Size of functional category

3 500

☐ Size of Q&T

2

☐ Numeric IDs treated as

MIM_GENE_ACC

☐ Significance threshold

Benjamini-Hochberg FDR

☐ Statistical domain size

Only annotated genes

Download g:Profiler data as GMT:

ENSG_name

☐ **[?] Gene Ontology** ☒ Biological process ☐ Cellular component ☐ Molecular function

☒ Inferred from experiment [IDA, IPI, IMP, IGI, IEP]

☐ Direct assay [IDA] / Mutant phenotype [IMP]

☐ Genetic interaction [IGI] / Physical interaction [IPI]

☐ Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]

☐ Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic context [IGC]

☐ Biological aspect of ancestor [IBA] / Rapid divergence [IRD]

☐ Reviewed computational analysis [RCA] / Electronic annotation [IEA]

☐ No biological data [ND] / Not annotated [NA]


☐ Biological pathways ☒ KEGG ☒ Reactome

☐ Regulatory motifs in DNA ☐ TRANSFAC TFBS ☐ miRBase microRNAs

☐ CORUM protein complexes

☐ Human Phenotype Ontology (sequence homologs in other species)

☐ BioGRID protein-protein interaction



[g:GOSt Gene Group Functional Profiling](#)
[g:Cocoa Compact Compare of Annotations](#)
[g:Convert Gene ID Converter](#)
[g:Sorter Expression Similarity Search](#)
[g:Orth Orthology search](#)

[Welcome!](#)
[About](#)
[Contact](#)
[Beta](#)
[Archives](#)

1. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) NAR 35 W193-W200 [PDF]

1. Reimand, T. Arak, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2011 update) Nucleic Acids Research 2011; doi: 10.1093/nar/gkr378 [PDF]

[?] Organism

Homo sapiens

CA12
FAM171B
CELSR2
RFTN1
SOC52
IL1R1
NPTN
IL20
LXN

[?] Query (genes, proteins, probes, term)

g:Profiler! Clear

Example or random query

Options

☒ Significant only

☐ Ordered query

☒ No electronic GO annotations

☐ Chromosomal regions

☒ Hierarchical sorting

☒ Hierarchical filtering

Show all terms (no filtering)

[?] Output type

Graphical (PNG)

Hide advanced options

☐ Measure underrepresentation

☐ Gene list as a stat. background

☐ 1.00 User p-value

[?] Size of functional category

3 500

[?] Size of Q&T

2

[?] Numeric IDs treated as

MIM_GENE_ACC

[?] Significance threshold

Benjamini-Hochberg FDR

[?] Statistical domain size

Only annotated genes

Download g:Profiler data as GMT:
ENSG, name

☐ **[?] Gene Ontology** ☒ Biological process ☐ Cellular component ☐ Molecular function

☒ Inferred from experiment [IDA, IPI, IMP, IGT, IEP]

☒ Direct assay [IDA] / Mutant phenotype [IMP]

☒ Genetic interaction [IGI] / Physical interaction [IPI]

☒ Traceable author [TAS] / Non-traceable author [NAS] / Inferred by curator [IC]

☒ Expression pattern [IEP] / Sequence or structural similarity [ISS] / Genomic context

☒ Biological aspect of ancestor [IBA] / Rapid divergence [IRD]

☒ Reviewed computational analysis [RCA] / Electronic annotation [IEA]

☐ No biological data [ND] / Not annotated [NA]

☒ **Biological pathways** ☒ KEGG ☒ Reactome

☒ **[?] Regulatory motifs in DNA** ☐ TRANSFAC TFBS ☐ miRBase microRNAs

☐ **[?] CORUM protein complexes**

☐ **[?] Human Phenotype Ontology** (sequence homologs in other species)

☐ **[?] BioGRID protein-protein interaction**

>> g:Convert
Gene ID converter

>> g:Orth
Orthology Search

>> g:Sorter
Expression Similarity Search

>> g:Cocoa
Compact Compare of Annotations

>> Static URL
Come back later

Warning: Some gene identifiers are ambiguous. Resolve these manually?

EXERCISE 2: Step 9 (continued)

Warning: Some gene identifiers are ambiguous. Resolve these manually?

Attempt to automatically resolve symbols using a namespace (percentage of ambiguous symbols resolved in brackets):

- VEGA_GENE (12%)
- HGNC (12%)

ARHGAP8

- ☒ ENSG00000241484 (ARHGAP8, 9 GO annot.) - Rho GTPase activating protein 8 [Source:HGNC Symbol;Acc:HGNC:677]
- ☐ ENSG00000248405 (PRR5-ARHGAP8, 9 GO annot.) - PRR5-ARHGAP8 readthrough [Source:HGNC Symbol;Acc:HGNC:34512]
- ☐ Ignore this gene

BOLA2

- ☒ ENSG00000169627 (BOLA2B, 1 GO annot.) - bolA family member 2B [Source:HGNC Symbol;Acc:HGNC:32479]
- ☐ ENSG00000183336 (BOLA2, 1 GO annot.) - bolA family member 2 [Source:HGNC Symbol;Acc:HGNC:29488]
- ☐ Ignore this gene

GPR89B

- ☒ ENSG00000117262 (GPR89A, 14 GO annot.) - G protein-coupled receptor 89A [Source:HGNC Symbol;Acc:HGNC:31984]
- ☐ ENSG00000188092 (GPR89B, 14 GO annot.) - G protein-coupled receptor 89B [Source:HGNC Symbol;Acc:HGNC:13840]
- ☐ Ignore this gene

MRPS17

- ☒ ENSG00000239789 (MRPS17, 12 GO annot.) - mitochondrial ribosomal protein S17 [Source:HGNC Symbol;Acc:HGNC:14047]
- ☐ ENSG00000249773 (MRPS17, 6 GO annot.) - 28S ribosomal protein S17, mitochondrial {ECO:0000313|Ensembl:ENSP00000390331}; HCG1984214, isoform CRA_a {ECO:0000313|E...
- ☐ Ignore this gene

PRICKLE4

- ☒ ENSG00000124593 (PRICKLE4, 4 GO annot.) - prickle homolog 4 (Drosophila) [Source:HGNC Symbol;Acc:HGNC:16805]
- ☐ ENSG00000278224 (PRICKLE4, 4 GO annot.) - prickle homolog 4 (Drosophila) [Source:HGNC Symbol;Acc:HGNC:16805]
- ☐ Ignore this gene

SERPINA3

- ☒ ENSG00000273259 (SERPINA3, 14 GO annot.) - serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 [Source:HGNC Symbol;Acc:HGNC:16]
- ☐ ENSG00000196136 (SERPINA3, 14 GO annot.) - serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3 [Source:HGNC Symbol;Acc:HGNC:16]
- ☐ Ignore this gene

SGK3

- ☒ ENSG00000104205 (SGK3, 23 GO annot.) - serum/glucocorticoid regulated kinase family, member 3 [Source:HGNC Symbol;Acc:HGNC:10812]
- ☐ ENSG00000270024 (C8orf44-SGK3, 23 GO annot.) - C8orf44-SGK3 readthrough [Source:HGNC Symbol;Acc:HGNC:48354]
- ☐ Ignore this gene

TXNDC5

- ☒ ENSG00000259040 (BLOC1S5-TXNDC5, 5 GO annot.) - BLOC1S5-TXNDC5 readthrough (NMD candidate) [Source:HGNC Symbol;Acc:HGNC:42001]
- ☐ ENSG00000239264 (TXNDC5, 14 GO annot.) - thioredoxin domain containing 5 (endoplasmic reticulum) [Source:HGNC Symbol;Acc:HGNC:21073]
- ☐ Ignore this gene

[Resubmit query](#)

EXERCISE 2: Step 10

source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Gene Ontology (Biological process)					
BP	negative regulation of cellular component organization	GO:0051129	385	392	22	3.44e-02
BP	regulation of microvillus assembly	GO:0032534	4	392	3	2.23e-02
BP	positive regulation of cell death	GO:0010942	423	392	23	5.00e-02
BP	regulation of protein kinase B signaling	GO:0051896	82	392	9	4.76e-02
BP	positive regulation of protein kinase B signaling	GO:0051897	56	392	9	2.06e-03
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (KEGG)					
ke	TGF-beta signaling pathway	KEGG:04350	79	386	8	5.00e-02
source	term name	term ID	n. of term genes	n. of query genes	n. of common genes	corrected p-value
	Biological pathways (Reactome)					
re	Regulation of mitotic cell cycle	REAC:453276	85	387	9	5.00e-02
re	APC/C-mediated degradation of cell cycle proteins	REAC:174143	85	387	9	5.00e-02