bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

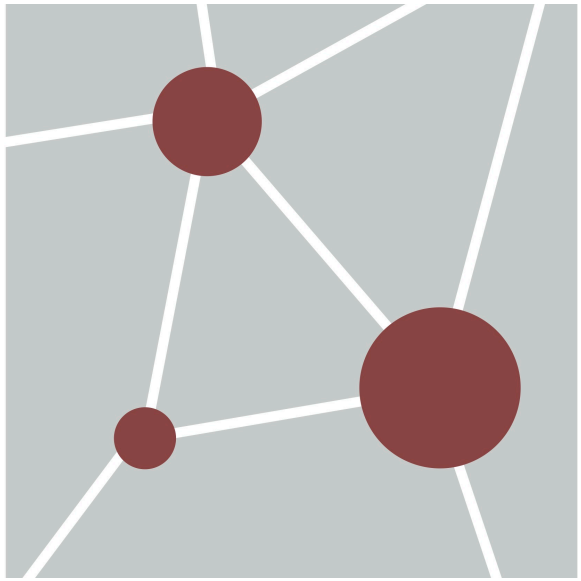
Finding over-represented pathways in gene lists: practical lab

Ruth Isserlin

Pathway and Network Analysis

June 5-7, 2023

bioinformatics.ca



BADER LAB

UNIVERSITY OF TORONTO

Donnelly Centre
Cellular & Biomolecular Research
UNIVERSITY OF TORONTO

Learning Objectives of Module

- By the end of this lab, you will:
  - Be able to run a simple enrichment tool like **g:Profiler** using a gene list and understand the main parameters and output results.
  - Be able to run **GSEA** (Gene Set Enrichment Tool) on a ranked gene list and understand the main parameters and output results.

Part 1: g:Profiler

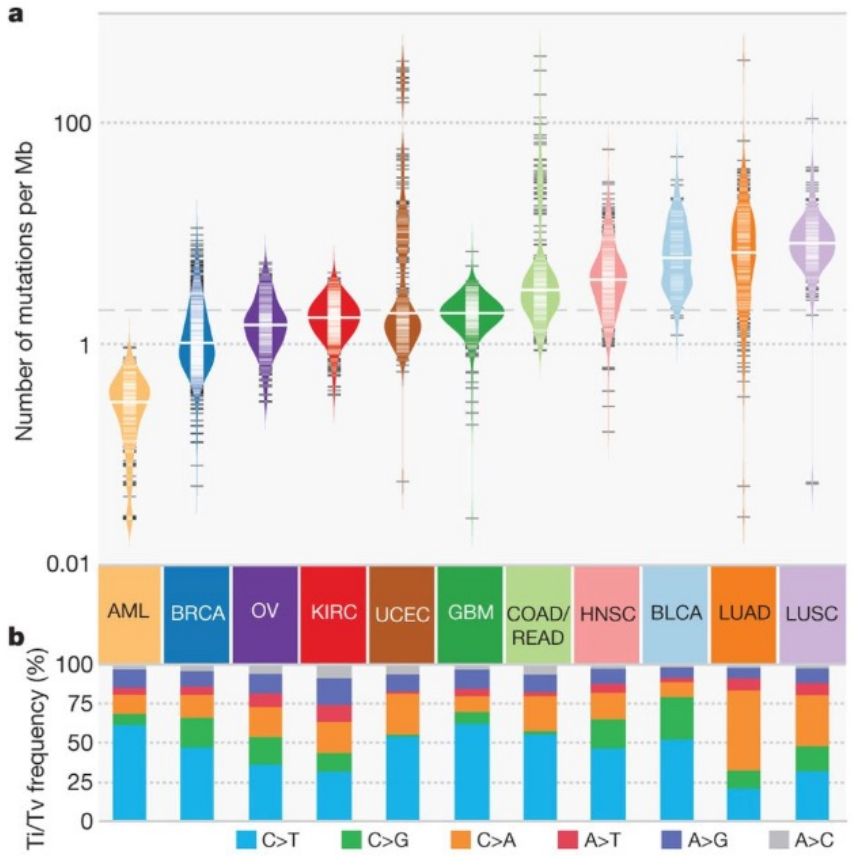
Part 2: GSEA Gene Set Enrichment Analysis

Characteristics:	g:Profiler	GSEA
Input	gene list (thresholded)	ranked gene list (non thresholded)
Statistics	**Fisher's exact test** (can upload specific background), **minimum hypergeometric test**	modified **Kolmogorov-Smirnov** test
Multiple hypothesis testing correction	yes (FDR, Bonferroni, custom)	yes (FDR)
Pathway databases (gene-sets) (choice/ up to date?)	several databases, can check the ones we are interested in, frequently updated	Several choices from MSigDB from GSEA or upload custom ones. link to Baderlab gene-sets both frequently updated
Model organisms	multiple, directly from Ensembl	mostly human through MSigDB but compatible with any model organisms using the custom upload function.
Output	Graphic image or table and compatible with Cytoscape/EnrichmentMap	Table and Compatible with Cytoscape/EnrichmentMap
Software type	Website and R package	Standalone (java) / or can be called and run from command line

Part 1:

Data used for practical lab:

Dataset: Mutational landscape and significance across 12 major cancer types



https://www.nature.com/articles/nature12634 (2013)

Exome sequencing
Tumor samples and
matched control tissues

↓

Detection of points mutations and small insertions/deletions: somatic variant calls in each cancer type and in each tumor

↓

Calculation of mutation frequency: genes mutated in at least 5% of tumors were selected

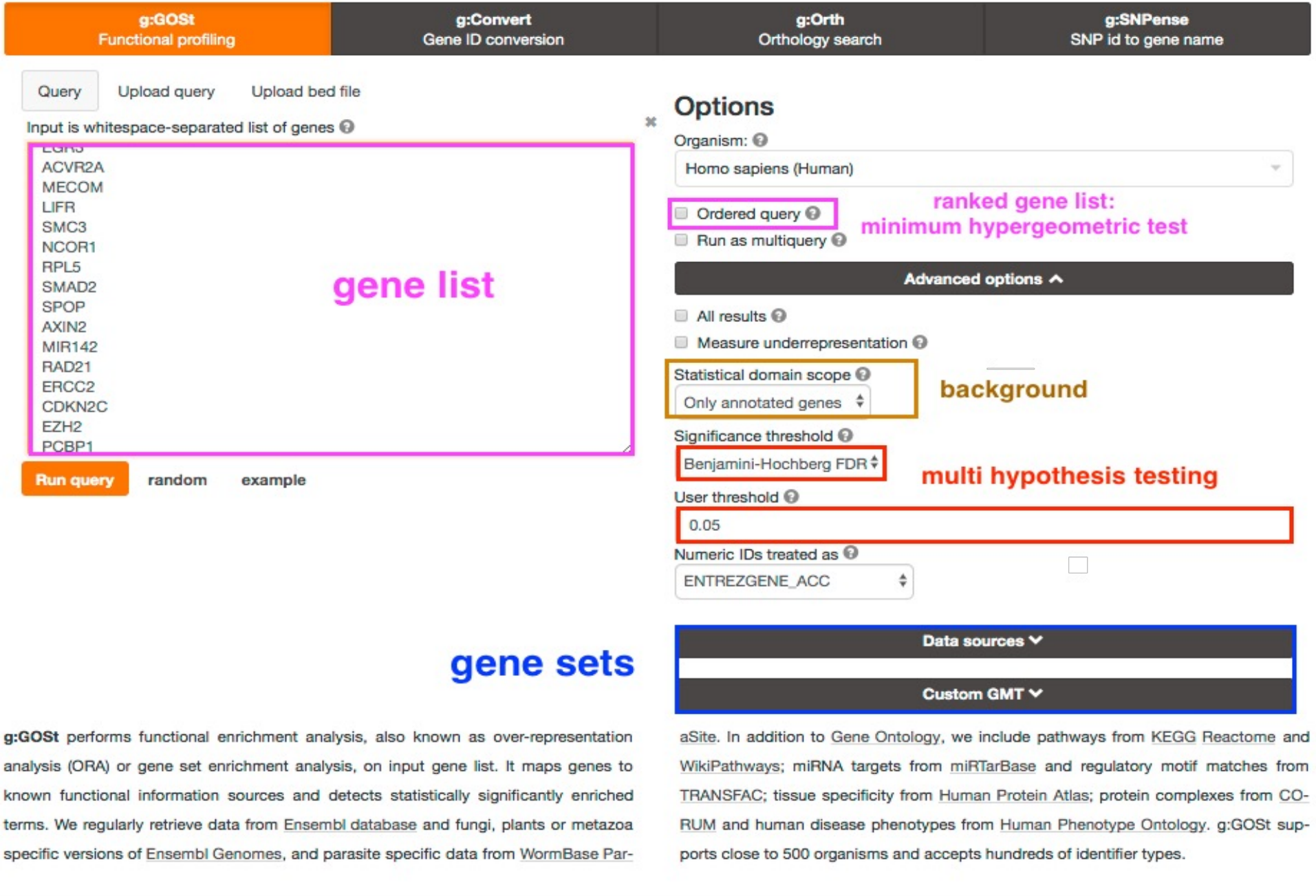
↓

Genes positively correlated with number of mutation per sample

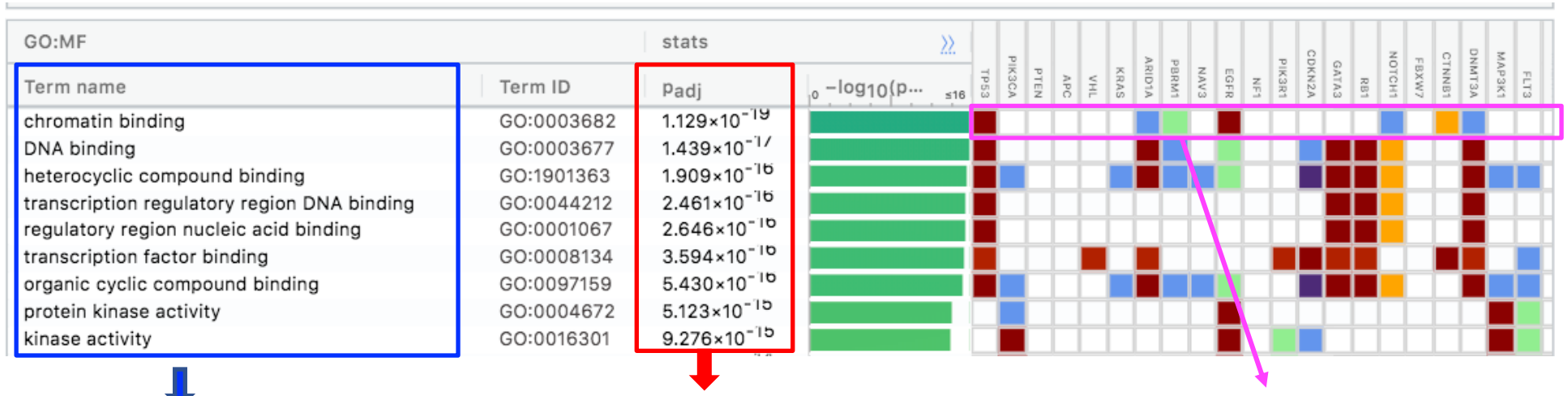
↓

127 'significantly mutated genes'

gene list


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
ACVR2A
MECOM
LIFR
SMC3
NCOR1
RPL5
SMAD2
SPOP
AXIN2
MIR142
RAD21
ERCC2
CDKN2C
EZH2
PCBP1
gene list
Run query
random
example
Options
Organism:
Homo sapiens (Human)
Ordered query
Run as multiquery
ranked gene list:
minimum hypergeometric test
Advanced options
All results
Measure underrepresentation
Statistical domain scope
Only annotated genes
background
Significance threshold
Benjamini-Hochberg FDR
multi hypothesis testing
User threshold
0.05
Numeric IDs treated as
ENTREZGENE_ACC
Data sources
Custom GMT
gene sets
g:GOSt performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. We regularly retrieve data from Ensembl database and fungi, plants or metazoa specific versions of Ensembl Genomes, and parasite specific data from WormBase ParaSite. In addition to Gene Ontology, we include pathways from KEGG Reactome and WikiPathways; miRNA targets from miRTarBase and regulatory motif matches from TRANSFAC; tissue specificity from Human Protein Atlas; protein complexes from CORUM and human disease phenotypes from Human Phenotype Ontology. g:GOSt supports close to 500 organisms and accepts hundreds of identifier types.

Explore results



each row is a gene-set (pathway)

Result of Fisher's exact test + multiple hypothesis correction:
gene-sets (pathways) are significantly enriched at FDR <0.05 (scientific notation: 5×10^{-2})

colored boxes: genes in our gene list that overlap with the tested gene-set

Note: observe that same genes are included in several enriched gene-sets (pathways).

Time to start practical part:

g:Profiler

- Go the the CBW course page and go to module 2.
- Open the ‘Lab practical part 1 (g:Profiler)’ document.
- Download required files on your computer.
- Do the exercise at your own pace and ask teaching assistants for help or questions.

Bonus – Run g:Profiler programmatically from R

- See example code - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/run-gprofiler-from-r.html

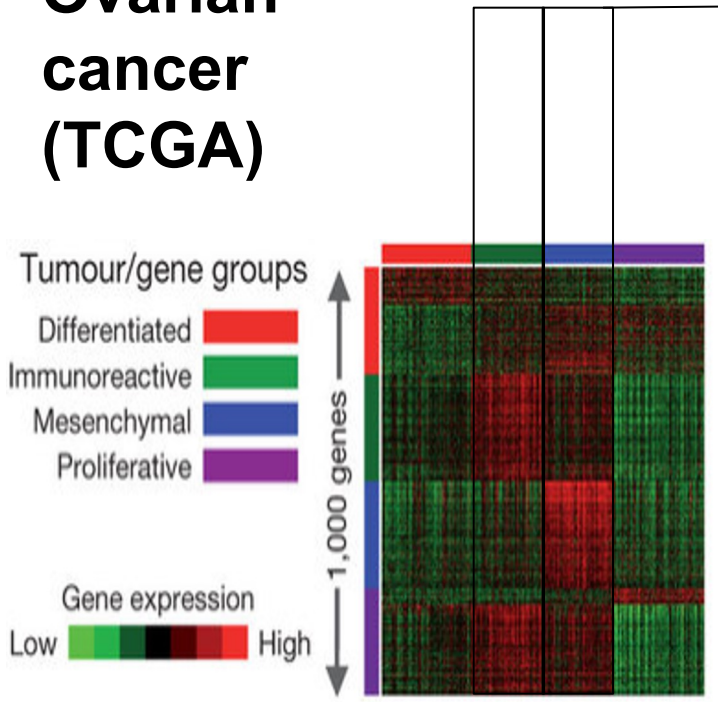
- For instructions on how to set up R so you can run the above notebooks - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/setup.html

Part 2:

Data used for practical lab: RNAseq workflow

Dataset

Ovarian cancer (TCGA)



RNAseq data

Immuno-reactive vs Mesen-chymal

Differential expression (edgeR)

Rank file

Gene-Set Enrichment Analysis (GSEA)

Integrated genomic analyses of ovarian carcinoma, PMID:21720365

Which files do we need to run GSEA?

- A ranked list of genes called the rank file
  - this is a text file (tab separated) that should be renamed to end with the extension .rnk
  - This file has 2 columns :
    - gene identifier
    - ranking values
- A file called a .gmt file that contains the pathway data base (the gene-sets)
  - this is a text file (tab separated) that should end with the extension .gmt
  - the first column contains gene-set names and the additional columns contains the gene names included in each gene-set

How to generate the rank file

genenames	logFC	logCPM	PValue	FDR
BGN	1.75	9.05	1.73E-33	2.50E-29
ANTXR1	1.55	7.50	4.39E-31	3.18E-27
FZD1	1.28	5.52	4.41E-30	2.13E-26
COL16A1	1.62	5.09	1.33E-29	4.81E-26
KLF3	0.13	6.37	8.32E-02	2.04E-01
RASEF	0.02	2.38	9.01E-01	9.49E-01
ISOC1	0.01	5.24	9.01E-01	9.50E-01
ANO1	0.03	4.93	9.02E-01	9.50E-01
CBWD3	-0.27	3.74	8.18E-02	2.02E-01
GBP4	-1.67	6.63	2.45E-16	2.57E-14
TAP1	-1.40	7.80	1.04E-19	2.38E-17
PSMB9	-1.55	6.52	1.84E-20	5.12E-18

edgeR output

1. Calculate the ranking score:

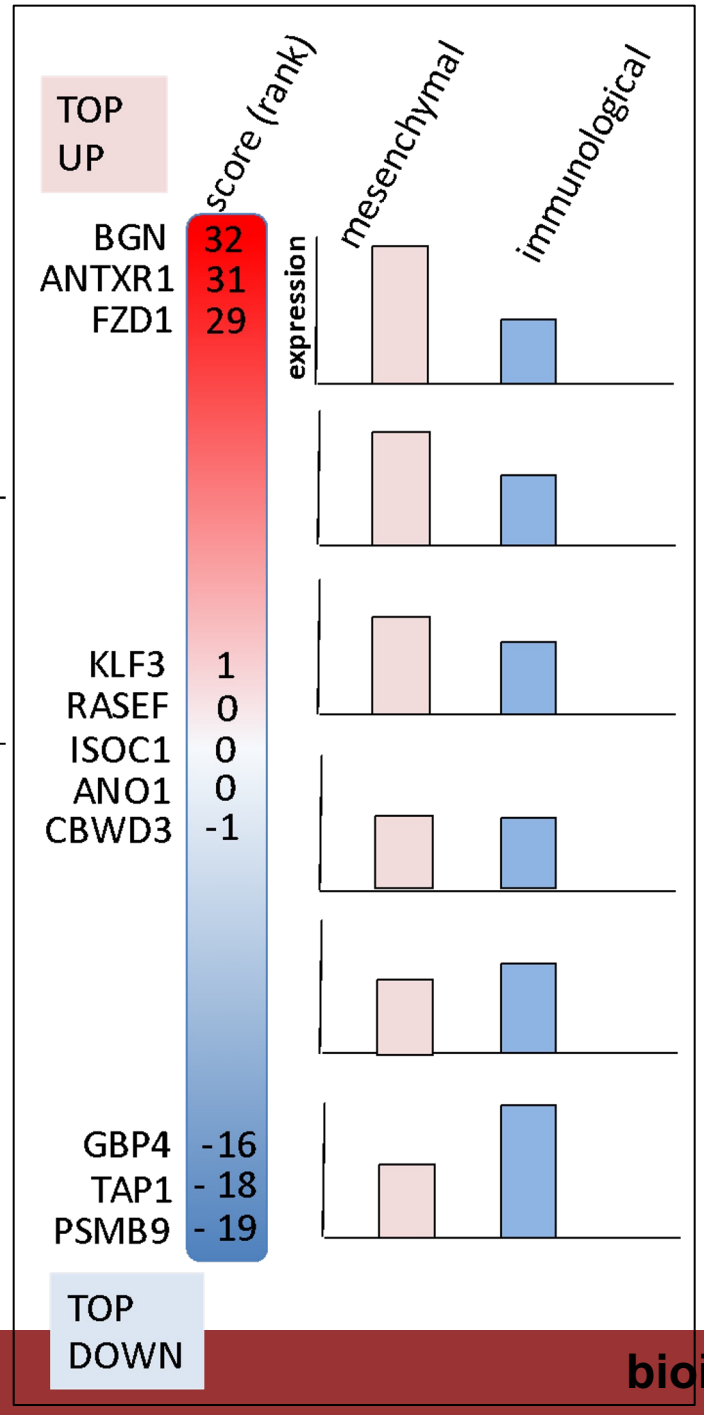
Using Excel:
=SIGN(logFC)*-LOG10(pvalue)

Using R:
sign(logFC)*-log10(pvalue)

gene name	score
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.61
TAP1	-18.98
PSMB9	-19.73

2. Save the file as a **tab** delimited text and with the extension **.rnk**

3.Do keep all genes in the rank files (e.g.15,000 genes) ! Do not remove non significant ones.



Ranked list (.rnk)

gene name	score
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.61
TAP1	-18.98
PSMB9	-19.73

Save the file as a **tab** delimited text and with the extension **.rnk**

Do keep all genes in the rank files (e.g.15,000 genes) ! Do not remove non significant ones.

What does a .gmt file look like?

Gene-set name	Gene-set name	gene	gene	gene	gene	gene	gene
MOLYBDENUM COFACTOR BIOSYNTHESIS%HUMANCYC%PWY-6823	molybdenum cofactor biosynthesis	NFS1	MOCS2	GPHN	MOCS3		
GLYCEROL DEGRADATION I%HUMANCYC%PWY-4261	glycerol degradation I	GK5	GK	GK2			
OXIDATIVE ETHANOL DEGRADATION III%HUMANCYC%PWY66-161	oxidative ethanol degradation III	CYP2E1	ACSS2	ACSS3	ALDH3A2	ACSS1	ALDH2
TETRAPYRROLE BIOSYNTHESIS II%HUMANCYC%PWY-5189	tetrapyrrole biosynthesis I	ALAS2	ALAD	UROS	HMBS	ALAS1	

* Save as tab delimited text with extension .gmt

Where to find a .gmt file?

If your model organism is Homo sapiens, you don't need to create your own:

- you can use directly the MSigDB within GSEA
- you can use the Baderlab gene-set file which is a frequently updated .gmt file which gathers public Gene Ontology and pathways from different sources.

If your model organism is Mus musculus:

- you can use the Baderlab gene-set file

If your model organism is different and you need to run GSEA:

- get (access or download) the Gene ontology database directly from biomart / Ensembl and parse it as a .gmt file (see coding example - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/create-gmt-file-from-ensembl.html).

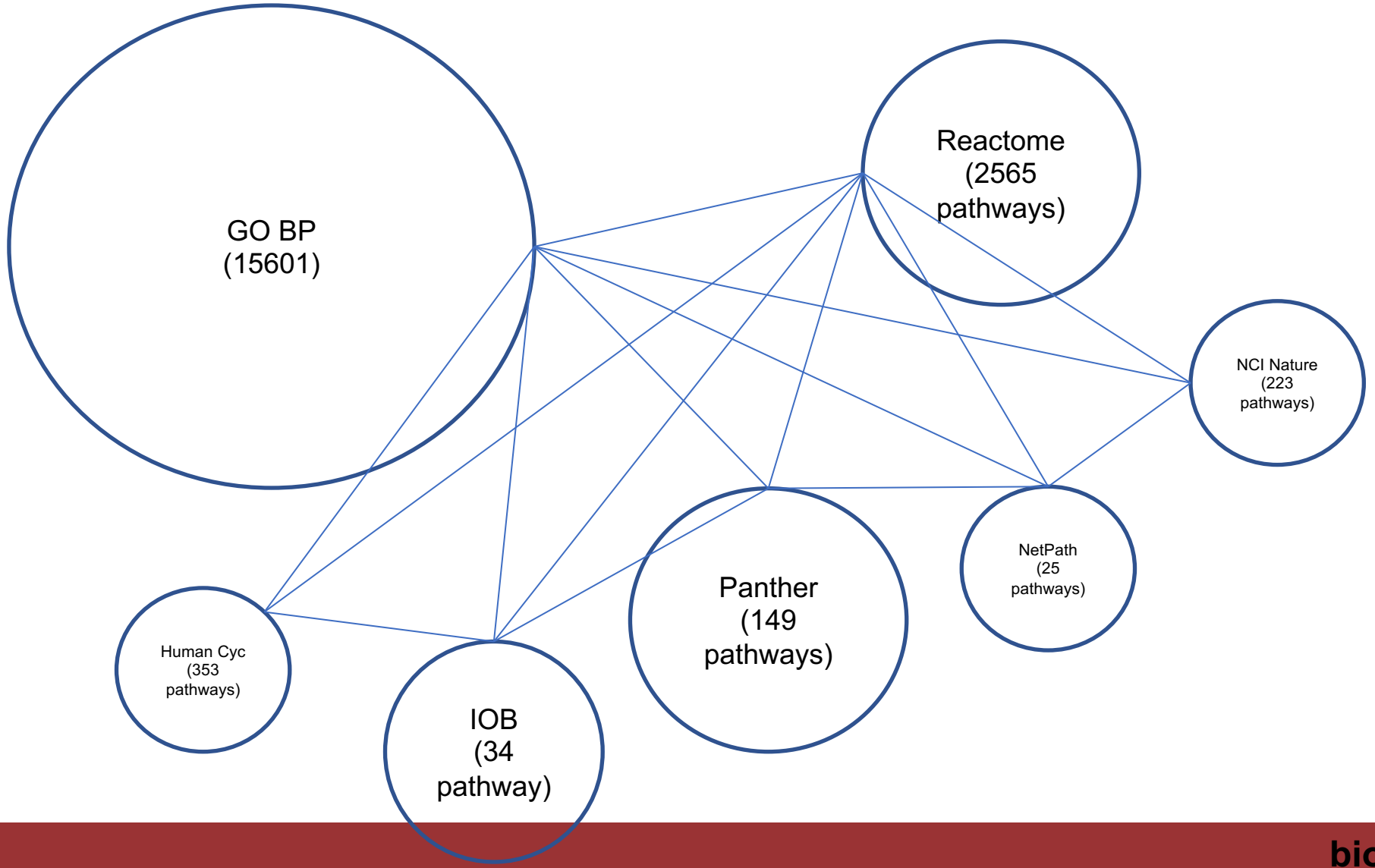
MSigDB database

https://www.gsea-msigdb.org/gsea/msigdb/

C2: curated gene sets (browse 6495 gene sets)	Gene sets in this collection are curated from various sources, including online pathway databases and the biomedical literature. Many sets are also contributed by individual domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into the following two subcollections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP). details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
Reactome subset of CP (browse 1654 gene sets)	Canonical Pathways gene sets derived from the Reactome pathway database.	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
C5: ontology gene sets (browse 15937 gene sets)	Gene sets that contain genes annotated by the same ontology term. The C5 collection is divided into two subcollections, the first derived from the Gene Ontology resource (GO) which contains BP, CC, and MF components and a second derived from the Human Phenotype Ontology (HPO). details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
BP: subset of GO (browse 7751 gene sets)	Gene sets derived from the GO Biological Process ontology.	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle
H: hallmark gene sets (browse 50 gene sets)	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. details	Download GMT Files Gene Symbols NCBI (Entrez) Gene IDs JSON bundle

BaderLab EM_Genesets

http://download.baderlab.org/EM_Genesets/



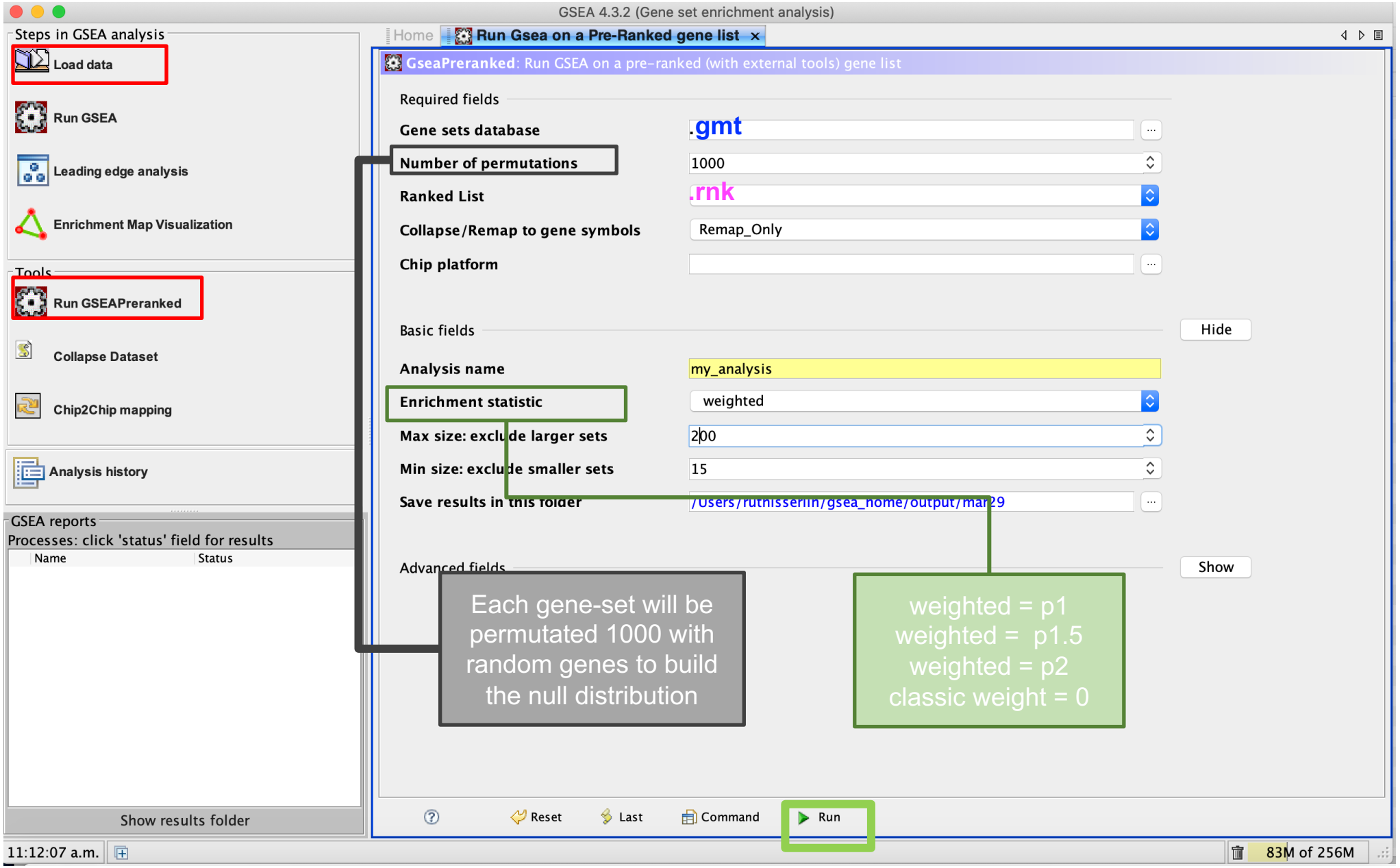
BaderLab EM_Genesets

- go to http://download.baderlab.org/EM_Genesets/
  - select current release/
    - Human/
      - symbol/
        - save the Human_GOPP_AllPathways_no_GO_iea….gmt file on your computer (right click on the link to save it)

Index of /EM_Genesets/current_release/Human/symbol

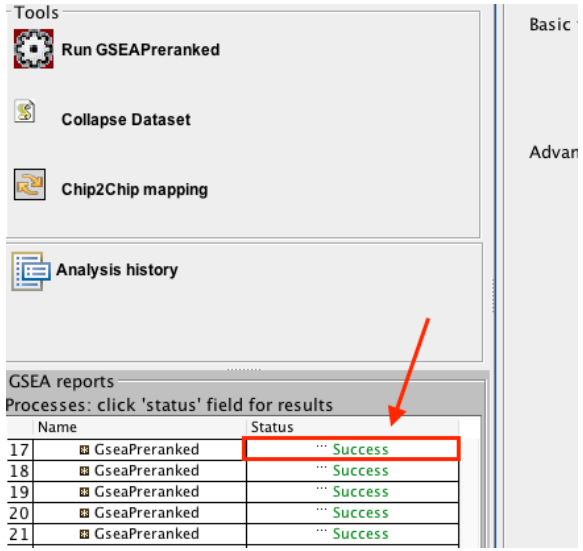
Name	Last modified	Size	Description
Parent Directory		-	
symbol_translation_summary.log	2023-04-03 17:18	465	
Human_GOBP_AllPathways_no_GO_iea_April_02_2023_symbol.gmt	2023-04-03 17:18	8.1M	
Human_GOBP_AllPathways_with_GO_iea_April_02_2023_symbol.gmt	2023-04-03 17:18	9.9M	
Human_GO_AllPathways_no_GO_iea_April_02_2023_symbol.gmt	2023-04-03 17:18	13M	
Human_GO_AllPathways_with_GO_iea_April_02_2023_symbol.gmt	2023-04-03 17:18	15M	
Human_AllPathways_April_02_2023_symbol.gmt	2023-04-03 17:18	1.7M	
Misc/	2023-04-03 17:18	-	
DrugTargets/	2023-04-03 17:18	-	
DiseasePhenotypes/	2023-04-03 17:18	-	
TranscriptionFactors/	2023-04-03 17:18	-	
miRs/	2023-04-03 17:18	-	
Pathways/	2023-04-03 17:18	-	
GO/	2023-04-03 17:18	-	

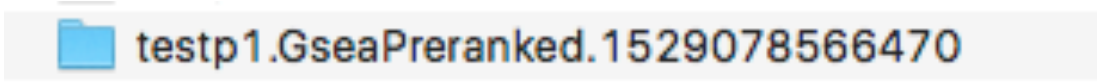
GSEA preranked



Exploring GSEA results

How to access GSEA results?





A GSEA result folder contains multiple files:

- **Index.html** will guide you to main result file
- The **edb folder** contains the input files filtered by GSEA
- **.rpt file** can be used in EnrichmentMap to built a network
- The main GSEA results are in 2 excel files :
  - **gsea_report**_for_pos_1401563306908.xls
  - **gsea_report**_for_neg_1401563306908.xls

GSEA Report for Dataset MesenchymalvsImmunoreactive_edger_ranks

Enrichment in phenotype: na Mesen

- 3173 / 5538 gene sets are upregulated in phenotype **na_pos**
- 1268 gene sets are significant at FDR < 25%
- 695 gene sets are significantly enriched at nominal pvalue < 1%
- 1037 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

gene-sets enriched in genes up-regulated in treated cells compared to non-treated samples or condition A vs condition B

Enrichment in phenotype: na Immuno

- 2365 / 5538 gene sets are upregulated in phenotype **na_neg**
- 1201 gene sets are significantly enriched at FDR < 25%
- 654 gene sets are significantly enriched at nominal pvalue < 1%
- 943 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in TSV format (tab delimited text)
- Guide to interpret results

gene-sets enriched in genes down-regulated in treated cells compared to non-treated samples or condition B vs condition A

Dataset details

- The dataset has 14449 features (genes)
- No probe set => gene symbol collapsing was requested, so all 14449 features were used

Gene set details

- Gene set size filters (min=15, max=200) resulted in filtering out 14170 / 19708 gene sets
- The remaining 5538 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 na_pos *versus* na_neg comparison

- The dataset has 14449 features (genes)
- Detailed rank ordered gene list for all features in the dataset

Index.html
summary of results

- Give the number or significant gene-sets (pathways)
- Link to the GSEA plots (snapshots)
- Link to the GSEA results as tabular format (html or excel format)

Note: you can access the index.html file using the **'Success'** link or locate it in the GSEA folder result.

Exploring GSEA Results

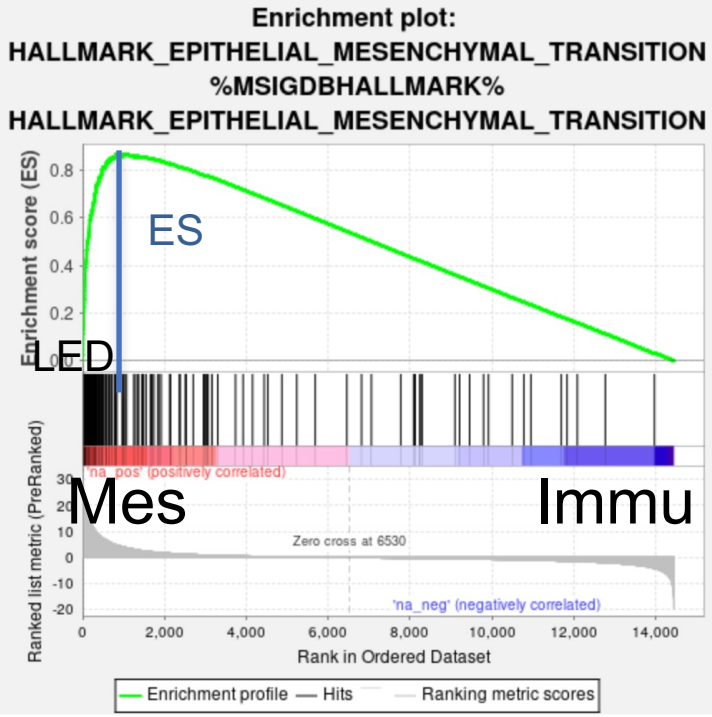
NES FDR

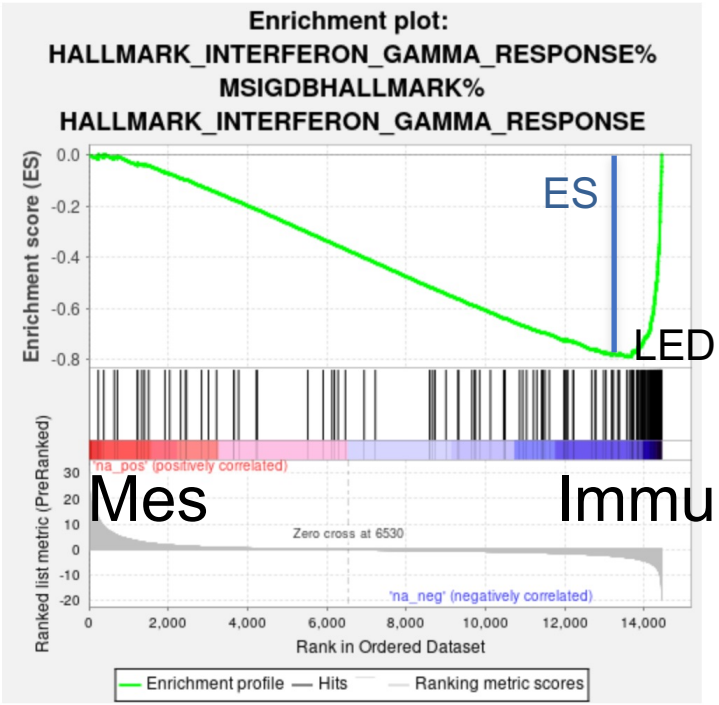
	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION%MSIGDBHALLMARK%HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	Details ...	145	0.87	2.39	0.000	0.000	0.000	1074	tags=58%, list=7%, signal=62%
2	EXTERNAL ENCAPSULATING STRUCTURE ORGANIZATION%GOBP%GO:0045229	Details ...	150	0.84	2.32	0.000	0.000	0.000	1123	tags=52%, list=8%, signal=56%
3	EXTRACELLULAR STRUCTURE ORGANIZATION%GOBP%GO:0043062	Details ...	149	0.84	2.31	0.000	0.000	0.000	1123	tags=52%, list=8%, signal=56%
4	EXTRACELLULAR MATRIX ORGANIZATION%GOBP%GO:0030198	Details ...	149	0.84	2.30	0.000	0.000	0.000	1123	tags=52%, list=8%, signal=56%
5	COLLAGEN FORMATION%REACTOME DATABASE ID RELEASE 83%1474290	Details ...	69	0.87	2.22	0.000	0.000	0.000	1123	tags=61%, list=8%, signal=66%
6	COLLAGEN BIOSYNTHESIS AND MODIFYING ENZYMES%REACTOME%R-HSA-1650814.4	Details ...	51	0.90	2.21	0.000	0.000	0.000	1002	tags=69%, list=7%, signal=73%
7	BETA1 INTEGRIN CELL SURFACE INTERACTIONS%PATHWAY INTERACTION DATABASE NCI-NATURE CURATED DATA%BETA1 INTEGRIN CELL SURFACE INTERACTIONS	Details ...	59	0.87	2.19	0.000	0.000	0.000	1283	tags=64%, list=9%, signal=70%

NES: normalized enrichment score
FDR: false discovery rate

Excel tables are going to be exported and uploaded in Cytoscape/EM (module 3)

Exploring GSEA Results





NES:2.40
FDR:0.0005

NES:-2.89
FDR: 0.0005

ES: enrichment score; NES: normalized enrichment score;
LED: leading edge genes; FDR false discovery rate

Time to start practical part:

GSEA
Gene Set Enrichment Analysis

- Go the the CBW course page.
- Download or open the Module 2 Lab practical documents.
- Download required files on your computer.
- Do the exercise at your own pace and ask teaching assistant for help or questions.

Links to more tutorials

Step by Step Protocol: Pathway enrichment analysis of -omics data:
https://www.nature.com/articles/s41596-018-0103-9

Notebooks of the protocol:
https://github.com/BaderLab/Cytoscape_workflows/tree/master/EnrichmentMapPipeline

https://baderlab.github.io/Cytoscape_workflows/EnrichmentMapPipeline/index.html

Bonus – Run GSEA programmatically from R

- See example code - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/run-gsea-from-within-r.ht

- For instructions on how to set up R so you can run the above notebooks - https://risserlin.github.io/CBW_pathways_workshop_R_notebooks/setup.html

We are on a Coffee Break & Networking Session

Workshop Sponsors: