The Gene Set Enrichment Analysis (GSEA)

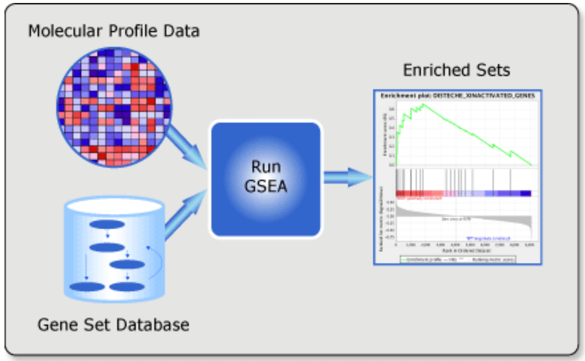
Rossella De Cegli, PhD

(TIGEM BAD days)

Monday July 17th 2023

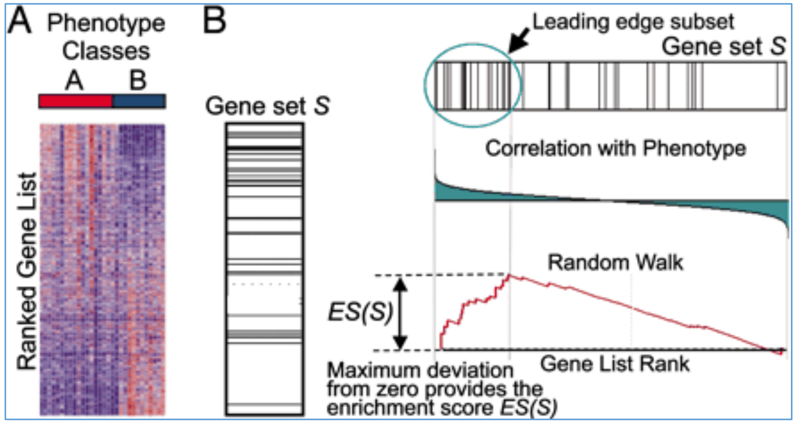
FONDAZIONE Telethon

- GSEA is a computational method that determines whether an a priori defined set of genes shows **statistically significant, concordant differences** between two biological states (phenotypes, conditions, treatments).



✧ GSEA considers experiments with genome-wide expression profiles from samples belonging to two classes, (**TREATED vs UNTREATED, KO vs WT** etc..).

✧ Genes are ranked based on the correlation between their expression and the class distinction by using any suitable metric

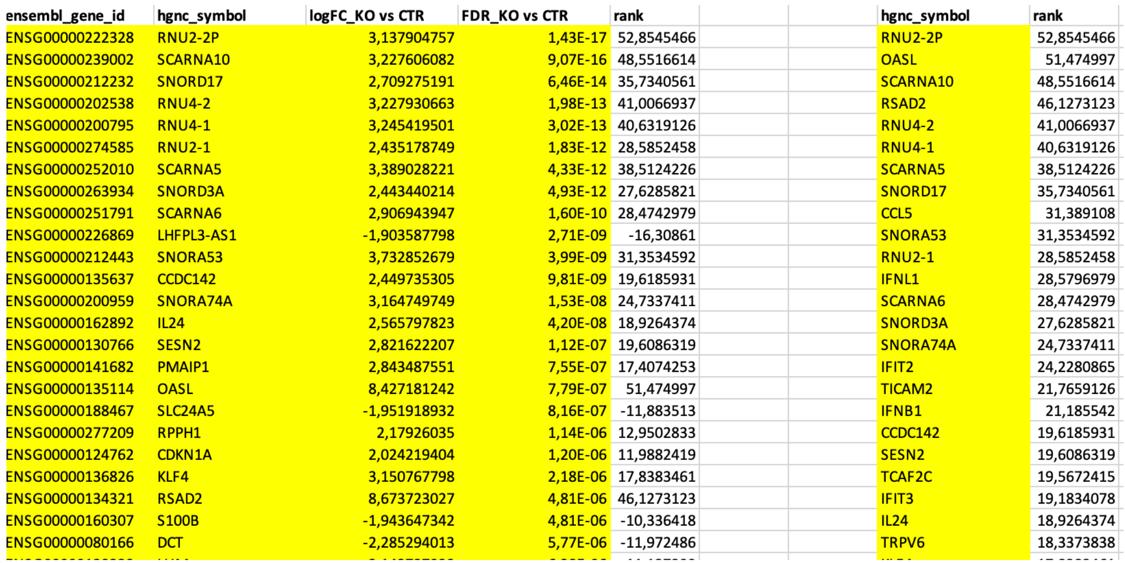


https://www.gsea-msigdb.org/gsea/msigdb/index.jsp

I Step:

Generate your ranked list of interest (3’DGE…Bulk_RNA-seq…. proteomics)

- The excel file will be converted in a **.rnk file**
- The complete gene list has to be in a format of two columns: one gene in one row
- To calculate the **rank**: use the formula =-LOG10(FDR)*logFC and rank the signed ratio (array) or the LogFC (RNA-seq/quantseq) from the top up SIGNIFICANT (FDR<5%) to the top dw values



ensembl_gene_id	hgnc_symbol	logFC_KO vs CTR	FDR_KO vs CTR	rank
ENSG00000222328	RNU2-2P	3,137904757	1,43E-17	52,8545466
ENSG00000239002	SCARNA10	3,227606082	9,07E-16	48,5516614
ENSG00000212232	SNORD17	2,709275191	6,46E-14	35,7340561
ENSG00000202538	RNU4-2	3,227930663	1,98E-13	41,0066937
ENSG00000200795	RNU4-1	3,245419501	3,02E-13	40,6319126
ENSG00000274585	RNU2-1	2,435178749	1,83E-12	28,5852458
ENSG00000252010	SCARNA5	3,389028221	4,33E-12	38,5124226
ENSG00000263934	SNORD3A	2,443440214	4,93E-12	27,6285821
ENSG00000251791	SCARNA6	2,906943947	1,60E-10	28,4742979
ENSG00000226869	LHFPL3-AS1	-1,903587798	2,71E-09	-16,30861
ENSG00000212443	SNORA53	3,732852679	3,99E-09	31,3534592
ENSG00000135637	CCDC142	2,449735305	9,81E-09	19,6185931
ENSG00000200959	SNORA74A	3,164749749	1,53E-08	24,7337411
ENSG00000162892	IL24	2,565797823	4,20E-08	18,9264374
ENSG00000130766	SESN2	2,821622207	1,12E-07	19,6086319
ENSG00000141682	PMAIP1	2,843487551	7,55E-07	17,4074253
ENSG00000135114	OASL	8,427181242	7,79E-07	51,474997
ENSG00000188467	SLC24A5	-1,951918932	8,16E-07	-11,883513
ENSG00000277209	RPPH1	2,17926035	1,14E-06	12,9502833
ENSG00000124762	CDKN1A	2,024219404	1,20E-06	11,9882419
ENSG00000136826	KLF4	3,150767798	2,18E-06	17,8383461
ENSG00000134321	RSAD2	8,673723027	4,81E-06	46,1273123
ENSG00000160307	S100B	-1,943647342	4,81E-06	-10,336418
ENSG00000080166	DCT	-2,285294013	5,77E-06	-11,972486

hgnc_symbol	rank
RNU2-2P	52,8545466
OASL	51,474997
SCARNA10	48,5516614
RSAD2	46,1273123
RNU4-2	41,0066937
RNU4-1	40,6319126
SCARNA5	38,5124226
SNORD17	35,7340561
CCL5	31,389108
SNORA53	31,3534592
RNU2-1	28,5852458
IFNL1	28,5796979
SCARNA6	28,4742979
SNORD3A	27,6285821
SNORA74A	24,7337411
IFIT2	24,2280865
TICAM2	21,7659126
IFNB1	21,185542
CCDC142	19,6185931
SESN2	19,6086319
TCAF2C	19,5672415
IFIT3	19,1834078
IL24	18,9264374
TRPV6	18,3373838

- Copy the two columns by using the application **TextWrangler (BBEdit)** to obtain the .rnk file.

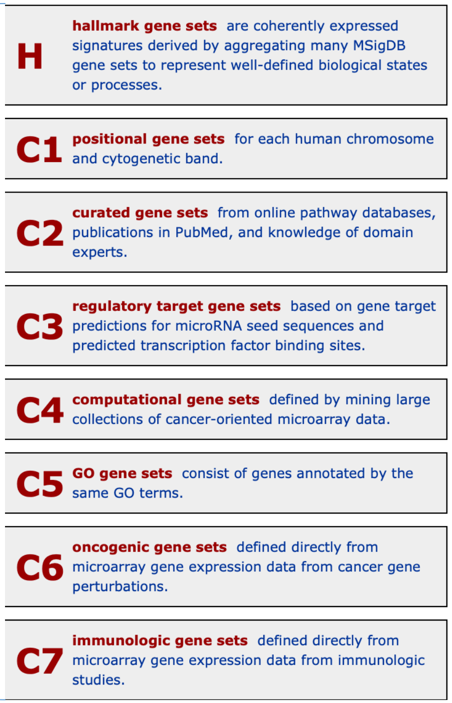
II Step A: Run a GSEA CUSTOM.

- **What is a Custom list:** it is a list of interest as a list of targets of a specific TF; a list of lysosomal genes, or of immune response-related genes…)
- This list may be generated as an excel file but to load the data this excel must be converted in a **.gmt file** (please use the application **TextWrangler**).
- The list has to be in a format of one row (past special as TRANSPOSE)
- In the first column of the obtained one row-list please write the NAME of your study

II StepB: Run a MsigDB GSEA

- The GSEA gene sets are divided into 8 major collections:

The Frequently used:



H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

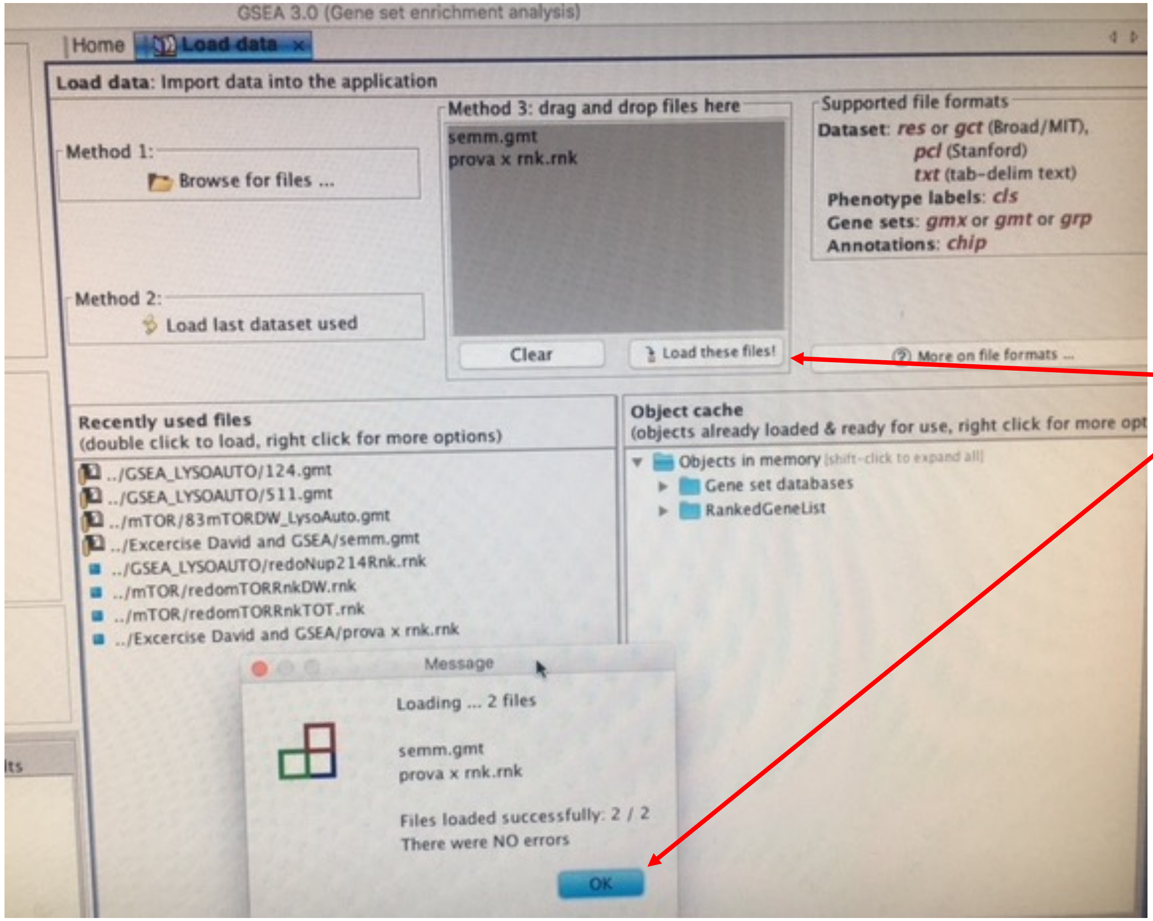
C3 **regulatory target gene sets** based on gene target predictions for microRNA seed sequences and predicted transcription factor binding sites.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.



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
semm.gmt
prova x rnk.rnk

Clear
Load these files!

Supported file formats
Dataset: *res* or *gct* (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)
Phenotype labels: *cls*
Gene sets: *gmx* or *gmt* or *grp*
Annotations: *chip*

More on file formats ...

Recently used files
(double click to load, right click for more options)
- ../GSEA_LYSOAUTO/124.gmt
- ../GSEA_LYSOAUTO/511.gmt
- ../mTOR/83mTORDW_LysoAuto.gmt
- ../Excercise David and GSEA/semm.gmt
- ../GSEA_LYSOAUTO/redoNup214Rnk.rnk
- ../mTOR/redomTORRnkDW.rnk
- ../mTOR/redomTORRnkTOT.rnk
- ../Excercise David and GSEA/prova x rnk.rnk

Object cache
(objects already loaded & ready for use, right click for more opt
- Objects in memory (shift-click to expand all)
  - Gene set databases
  - RankedGeneList

Message

Loading ... 2 files

semm.gmt
prova x rnk.rnk

Files loaded successfully: 2 / 2
There were NO errors

OK

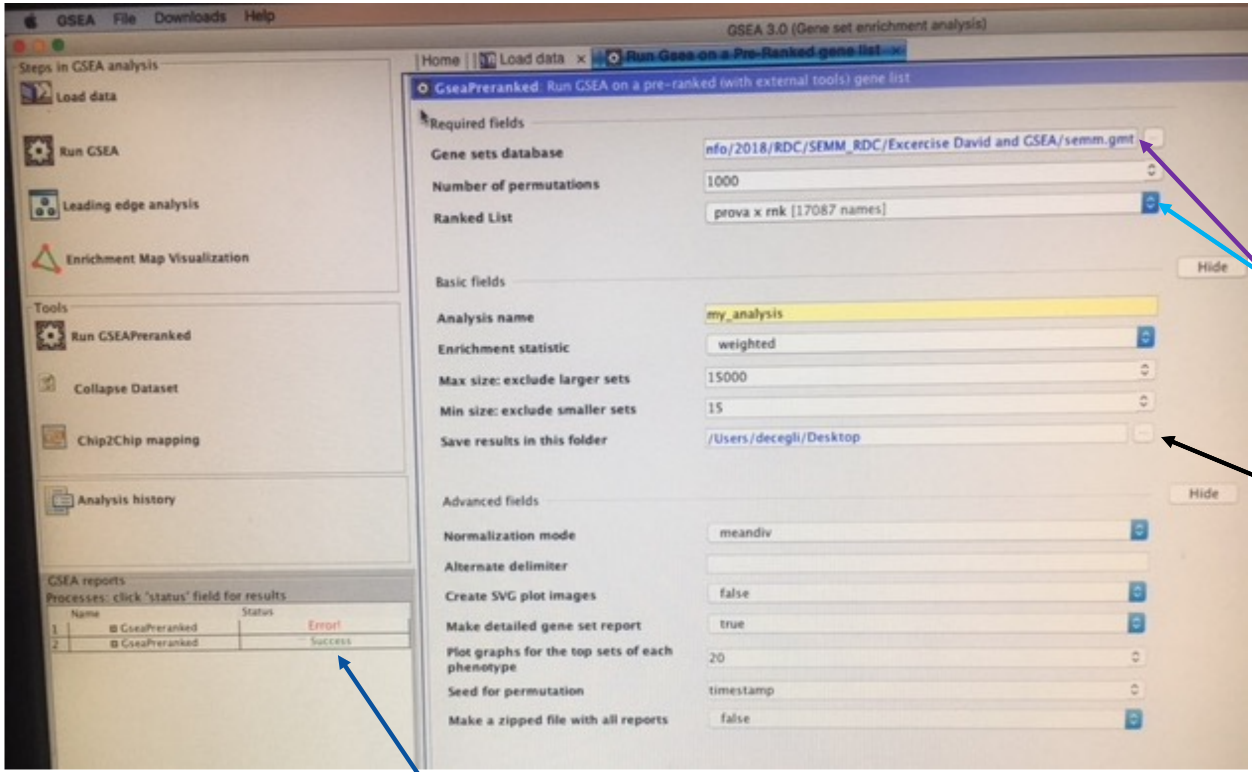
.rnk file
.gmt file
Successufully loaded

2. GSEA

Gene Set Enrichment Analysis

BROAD INSTITUTE

igem



GSEA 3.0 (Gene set enrichment analysis)

GSEA reports

Processes: click 'status' field for results

	Name	Status
1	GseaPreranked	Error!
2	GseaPreranked	Success

GseaPreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database	nfo/2018/RDC/SEMM_RDC/Excercise David and GSEA/semm.gmt
Number of permutations	1000
Ranked List	prova x rnk [17087 names]

Basic fields

Analysis name	my_analysis
Enrichment statistic	weighted
Max size: exclude larger sets	15000
Min size: exclude smaller sets	15
Save results in this folder	/Users/decegli/Desktop

Advanced fields

Normalization mode	meandiv
Alternate delimiter	
Create SVG plot images	false
Make detailed gene set report	true
Plot graphs for the top sets of each phenotype	20
Seed for permutation	timestamp
Make a zipped file with all reports	false

.rnk file

.gmt file

Where you wish to save the GSEA folder

GSEA
Success or
ERROR

Now ENJOY GSEA!!!!!

2. GSEA

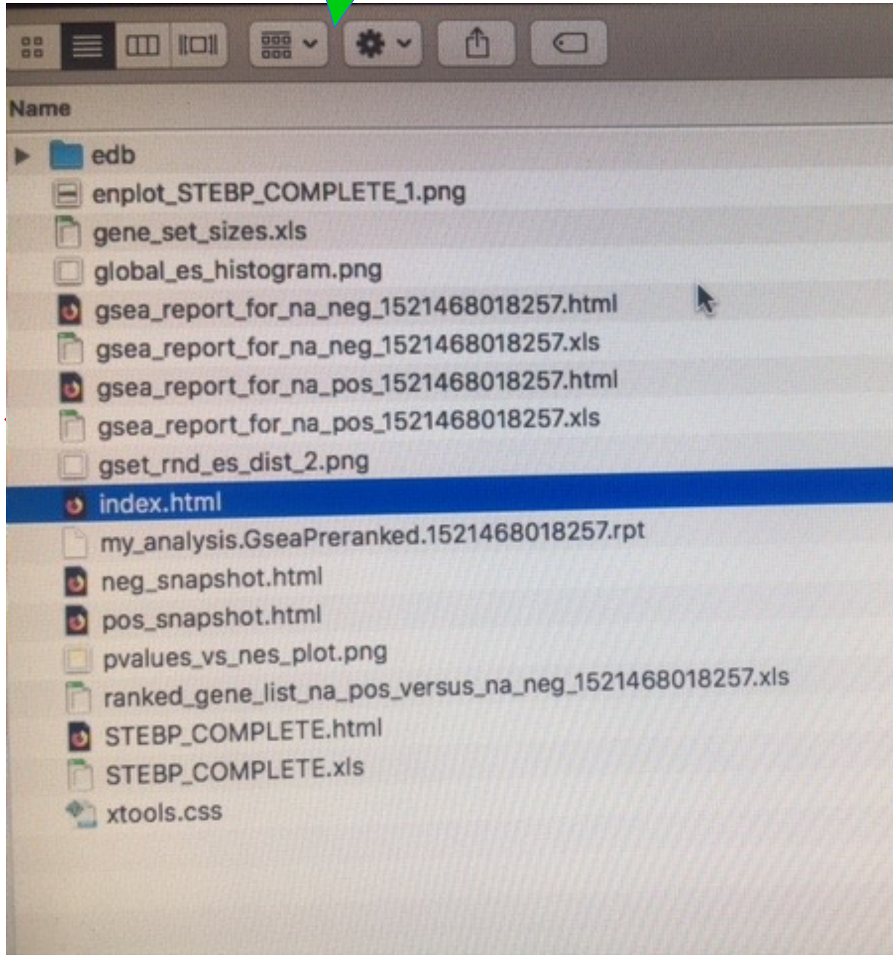
GSEA Gene Set Enrichment Analysis

BROAD INSTITUTE

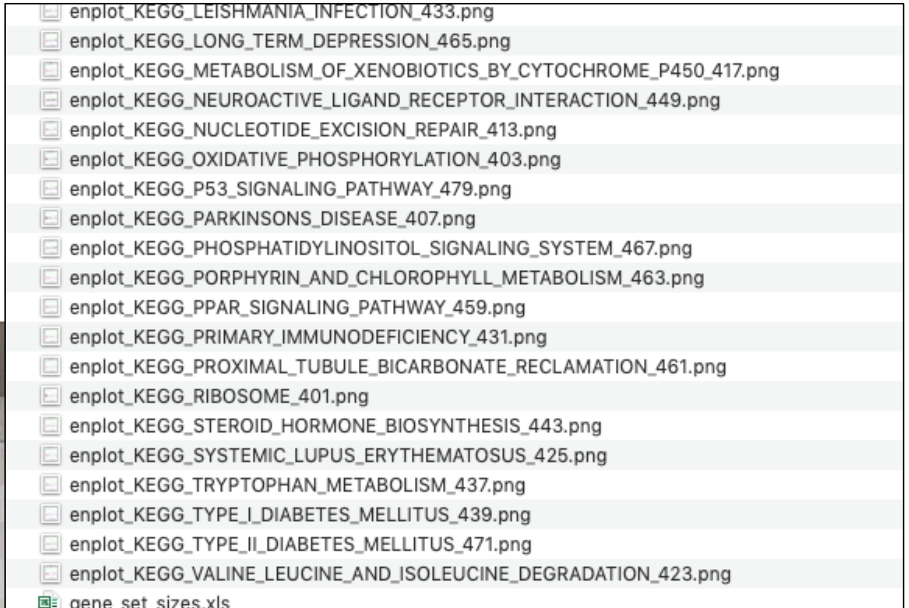
igem

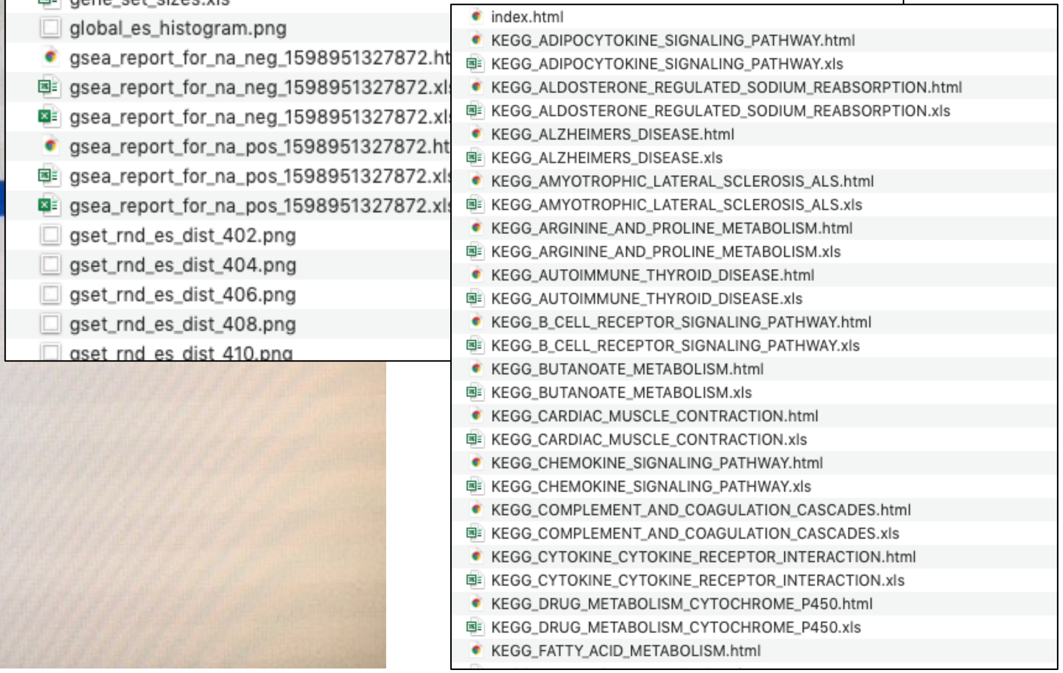
RESULTS

GSEA FOLDER



```
Name
▶ edb
enplot_STEBP_COMPLETE_1.png
gene_set_sizes.xls
global_es_histogram.png
gsea_report_for_na_neg_1521468018257.html
gsea_report_for_na_neg_1521468018257.xls
gsea_report_for_na_pos_1521468018257.html
gsea_report_for_na_pos_1521468018257.xls
gset_rnd_es_dist_2.png
index.html
my_analysis.GseaPreranked.1521468018257.rpt
neg_snapshot.html
pos_snapshot.html
pvalues_vs_nes_plot.png
ranked_gene_list_na_pos_versus_na_neg_1521468018257.xls
STEBP_COMPLETE.html
STEBP_COMPLETE.xls
xtools.css
```





```
enplot_KEGG_LEISHMANIA_INFECTION_433.png
enplot_KEGG_LONG_TERM_DEPRESSION_465.png
enplot_KEGG_METABOLISM_OF_XENOBIOTICS_BY_CYTOCHROME_P450_417.png
enplot_KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION_449.png
enplot_KEGG_NUCLEOTIDE_EXCISION_REPAIR_413.png
enplot_KEGG_OXIDATIVE_PHOSPHORYLATION_403.png
enplot_KEGG_P53_SIGNALING_PATHWAY_479.png
enplot_KEGG_PARKINSONS_DISEASE_407.png
enplot_KEGG_PHOSPHATIDYLINOSITOL_SIGNALING_SYSTEM_467.png
enplot_KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM_463.png
enplot_KEGG_PPAR_SIGNALING_PATHWAY_459.png
enplot_KEGG_PRIMARY_IMMUNODEFICIENCY_431.png
enplot_KEGG_PROXIMAL_TUBULE_BICARBONATE_RECLAMATION_461.png
enplot_KEGG_RIBOSOME_401.png
enplot_KEGG_STEROID_HORMONE_BIOSYNTHESIS_443.png
enplot_KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS_425.png
enplot_KEGG_TRYPTOPHAN_METABOLISM_437.png
enplot_KEGG_TYPE_I_DIABETES_MELLITUS_439.png
enplot_KEGG_TYPE_II_DIABETES_MELLITUS_471.png
enplot_KEGG_VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION_423.png
gene_set_sizes.xls
global_es_histogram.png
gsea_report_for_na_neg_1598951327872.ht
gsea_report_for_na_neg_1598951327872.xl
gsea_report_for_na_neg_1598951327872.xl
gsea_report_for_na_pos_1598951327872.ht
gsea_report_for_na_pos_1598951327872.xl
gsea_report_for_na_pos_1598951327872.xl
gset_rnd_es_dist_402.png
gset_rnd_es_dist_404.png
gset_rnd_es_dist_406.png
gset_rnd_es_dist_408.png
gset_rnd_es_dist_410.png
```

```
index.html
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY.html
KEGG_ADIPOCYTOKINE_SIGNALING_PATHWAY.xls
KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION.html
KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION.xls
KEGG_ALZHEIMERS_DISEASE.html
KEGG_ALZHEIMERS_DISEASE.xls
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS.html
KEGG_AMYOTROPHIC_LATERAL_SCLEROSIS_ALS.xls
KEGG_ARGININE_AND_PROLINE_METABOLISM.html
KEGG_ARGININE_AND_PROLINE_METABOLISM.xls
KEGG_AUTOIMMUNE_THYROID_DISEASE.html
KEGG_AUTOIMMUNE_THYROID_DISEASE.xls
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY.html
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY.xls
KEGG_BUTANOATE_METABOLISM.html
KEGG_BUTANOATE_METABOLISM.xls
KEGG_CARDIAC_MUSCLE_CONTRACTION.html
KEGG_CARDIAC_MUSCLE_CONTRACTION.xls
KEGG_CHEMOKINE_SIGNALING_PATHWAY.html
KEGG_CHEMOKINE_SIGNALING_PATHWAY.xls
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES.html
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES.xls
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION.html
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION.xls
KEGG_DRUG_METABOLISM_CYTOCHROME_P450.html
KEGG_DRUG_METABOLISM_CYTOCHROME_P450.xls
KEGG_FATTY_ACID_METABOLISM.html
```

2. GSEA

Table: GSEA Results Summary

Dataset	prova x rnk
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_pos
GeneSet	SREBP_CHIP
Enrichment Score (ES)	0.5765218
Normalized Enrichment Score (NES)	1.9980491
Nominal p-value	0.0
FDR q-value	0.0
FWER p-Value	0.0

Example of GOOD GSEA RESULT

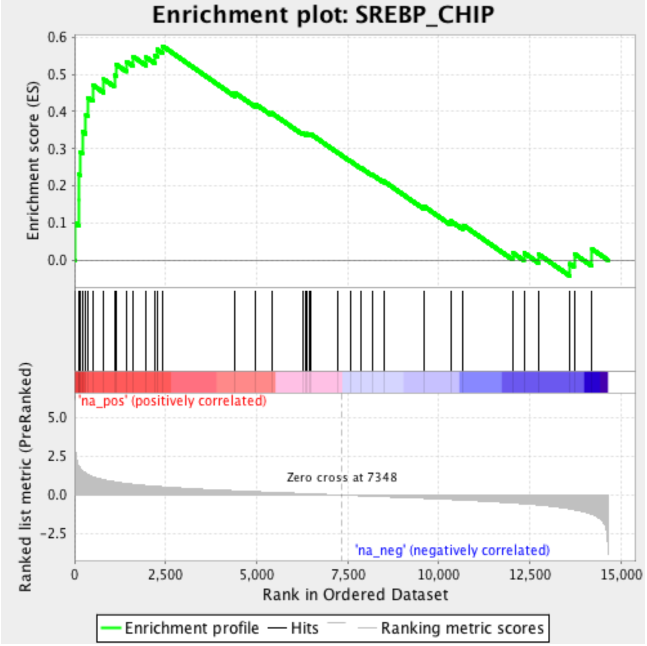


Fig 1: Enrichment plot: SREBP_CHIP
Profile of the Running ES Score & Positions of GeneSet Members on the Rank Ordered List

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
GO_LYSOSOMAL_LUMEN	69	-0,8664851	-1,9405853	0	0,00199854
GO_PIGMENT_GRANULE	96	-0,8154515	-1,8812673	0	0,00661115
GO_DEVELOPMENTAL_PIGMENTATION	39	-0,9059212	-1,8647331	0	0,00793784
GO_VACUOLAR_LUMEN	120	-0,7977715	-1,8842796	0	0,00844721
GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_ANTIGEN	151	-0,763359	-1,8666888	0	0,00918049
GO_ANTIGEN_PROCESSING_AND_PRESENTATION_OF_PEPTIDE_OR_POLYSACCHARIDE_ANTIGEN_VIA_MHC_CLASS_II	81	-0,8100114	-1,8460969	0	0,01338742
GO_VACUOLAR_PART	430	-0,6860185	-1,8328513	0	0,01725247
GO_AZUROPHIL_GRANULE	112	-0,749383	-1,8148923	0	0,02436469
GO_ANTIGEN_PROCESSING_AND_PRESENTATION	172	-0,741165	-1,8196093	0	0,02458064
GO_SPHINGOLIPID_METABOLIC_PROCESS	57	-0,8332759	-1,7893295	0	0,03715184

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
GO_RESPONSE_TO_TYPE_I_INTERFERON	68	0,9550428	1,7689795	0	0
GO_REGULATION_OF_VIRAL_GENOME_REPLICATION	74	0,94051254	1,7507216	0	0
GO_VIRAL_GENOME_REPLICATION	95	0,9257698	1,7638758	0	0
GO_DEFENSE_RESPONSE_TO_VIRUS	158	0,9159155	1,8068191	0	0
GO_RESPONSE_TO_VIRUS	216	0,9031988	1,8228018	0	0
GO_DEFENSE_RESPONSE_TO_OTHER_ORGANISM	238	0,88119084	1,7750893	0	0
GO_SM_LIKE_PROTEIN_FAMILY_COMPLEX	83	0,92373174	1,7440909	0	4,32E-04
GO_REGULATION_OF_VIRAL_LIFE_CYCLE	106	0,8948939	1,7309905	0	0,00202313
GO_NEGATIVE_REGULATION_OF_VIRAL_GENOME_REPLICATION	42	0,96347237	1,7121593	0	0,00567661
GO_SPLICEOSOMAL_SNRNP_ASSEMBLY	39	0,9559272	1,7155771	0	0,0059322
GO_NEGATIVE_REGULATION_OF_VIRAL_PROCESS	67	0,9221198	1,7126523	0	0,0061437
GO_REGULATION_OF_SYMBIOSIS_ENCOMPASSING_MUTUALISM_THROUGH_PARASITISM	169	0,85546035	1,7057966	0	0,00872576
GO_SNRNA_BINDING	39	0,9428287	1,7008914	0	0,0106801
GO_CYTOKINE_MEDIATED_SIGNALING_PATHWAY	476	0,80943793	1,6872447	0	0,02351773
GO_REGULATION_OF_T_CELL_MEDIATED_IMMUNITY	34	0,95067674	1,6772937	0,00166667	0,0383134
GO_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT	74	0,89211684	1,6725465	0	0,04628721
GO_DOUBLE_STRANDED_RNA_BINDING	66	0,89978564	1,653459	0	0,08593836
GO_REGULATION_OF_RECEPTOR_SIGNALING_PATHWAY_VIA_STAT	59	0,90091246	1,6586115	0,00154083	0,08643796

For Students: please refer to the file "GSEA_how create an input_example.xlsx", second sheet

- ✧ The threshold for statistical significance of GSEA is **FDR<0,25** and **Enrichment Score >0.5** for induced gene sets and Enrichment Score **<-0.5** for inhibited GS

2. GSEA

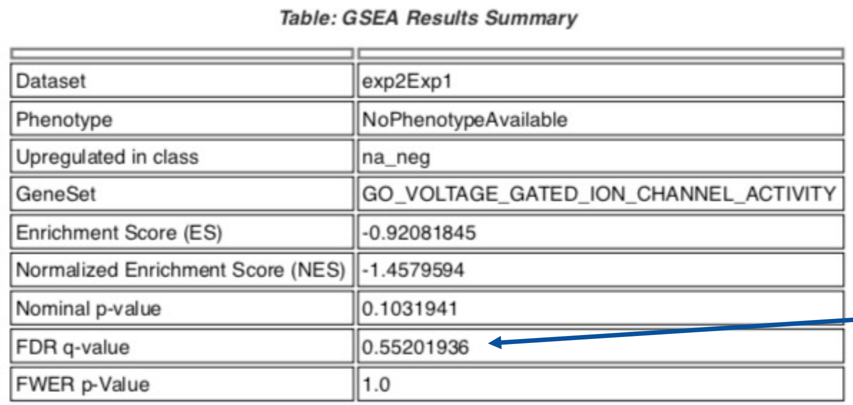
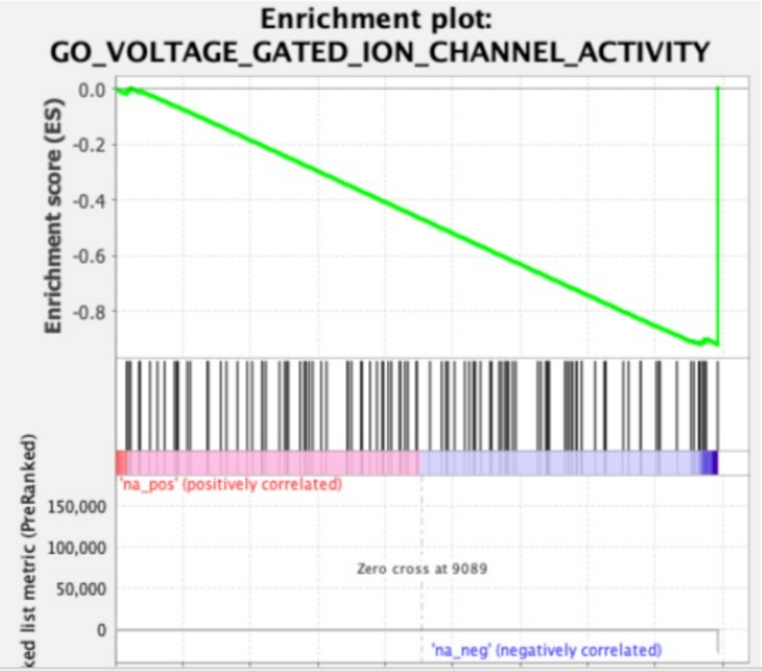


Table: GSEA Results Summary

Dataset	exp2Exp1
Phenotype	NoPhenotypeAvailable
Upregulated in class	na_neg
GeneSet	GO_VOLTAGE_GATED_ION_CHANNEL_ACTIVITY
Enrichment Score (ES)	-0.92081845
Normalized Enrichment Score (NES)	-1.4579594
Nominal p-value	0.1031941
FDR q-value	0.55201936
FWER p-Value	1.0

Example of NOT significant RESULT



✧ The threshold for statistical significance of GSEA is **FDR<0,25** and **Enrichment Score >0.5** for induced gene sets and Enrichment Score **<-0.5** for inhibited GS