Bioinformatics for Pathway Enrichment Analysis Part 2

Veronique Voisin and Ruth Isserlin

Computer Ontario Summer school 2025

Course outline

- General concept of pathway enrichment analysis with a ranked list.
- How do we perform pathway enrichment analysis (steps)?
- Example :TCGA pancreatic cancer RNASeq expression
- Practical lab: pathway enrichment analysis using R and Cytoscape

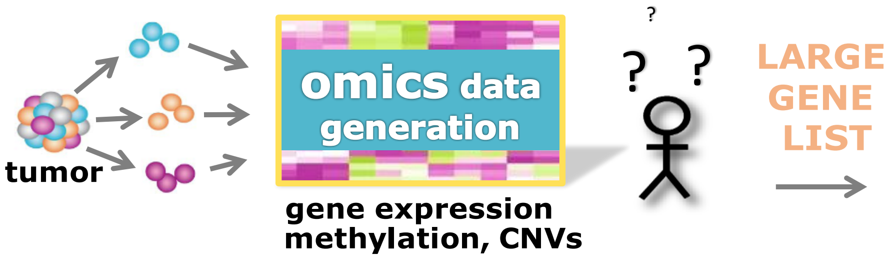
Learning Objectives

- Be able to understand:

  - Be able to run GSEA(Gene Set Enrichment Analysis) on a ranked gene list and understand the main parameters and output results.
  - Understand the different ways we can visualize enrichment results
  - the advantages of these different pathway enrichment analysis and visualizations.

General Workflow of pathway enrichment analysis

Step1



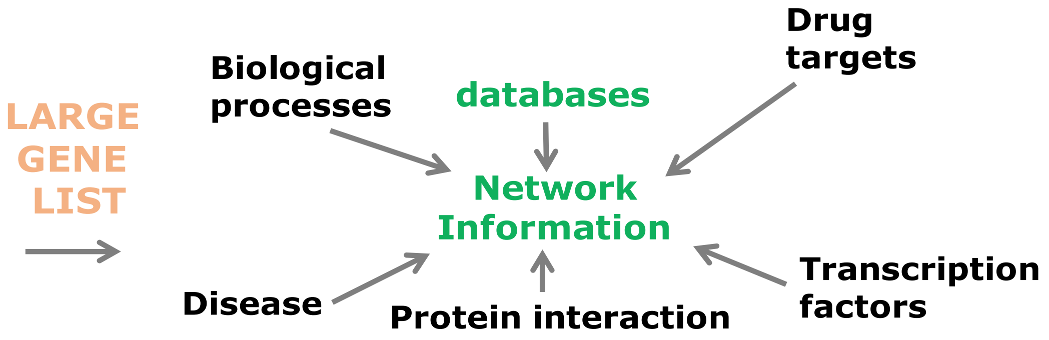
tumor

omics data generation

gene expression
methylation, CNVs

LARGE
GENE
LIST

Step2



LARGE
GENE
LIST

Biological processes

databases

Drug targets

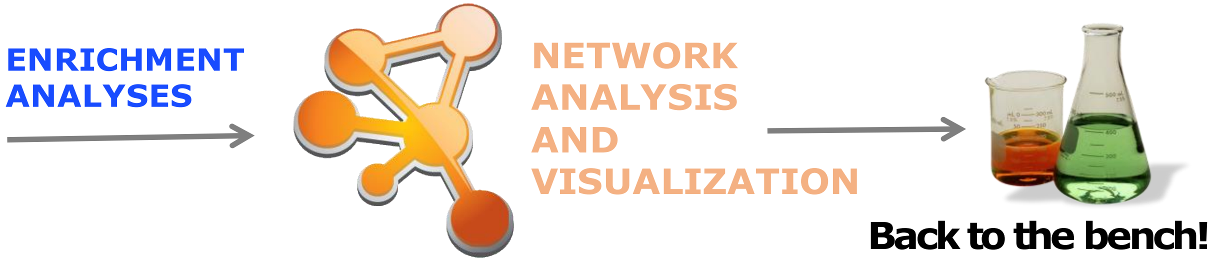
Network Information

Disease

Protein interaction

Transcription factors

Step3



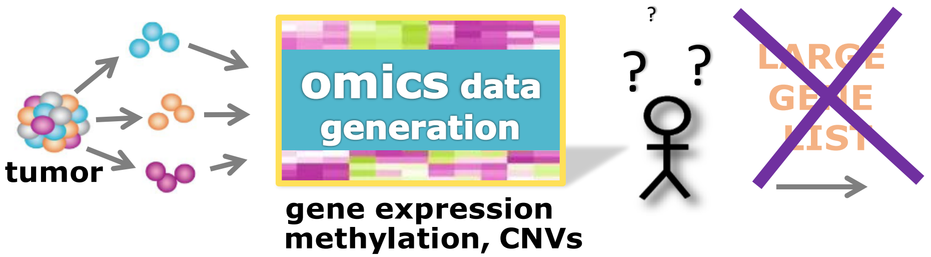
ENRICHMENT ANALYSES

NETWORK
ANALYSIS
AND
VISUALIZATION

Back to the bench!

General Workflow of pathway enrichment analysis

Step1



tumor

omics data generation

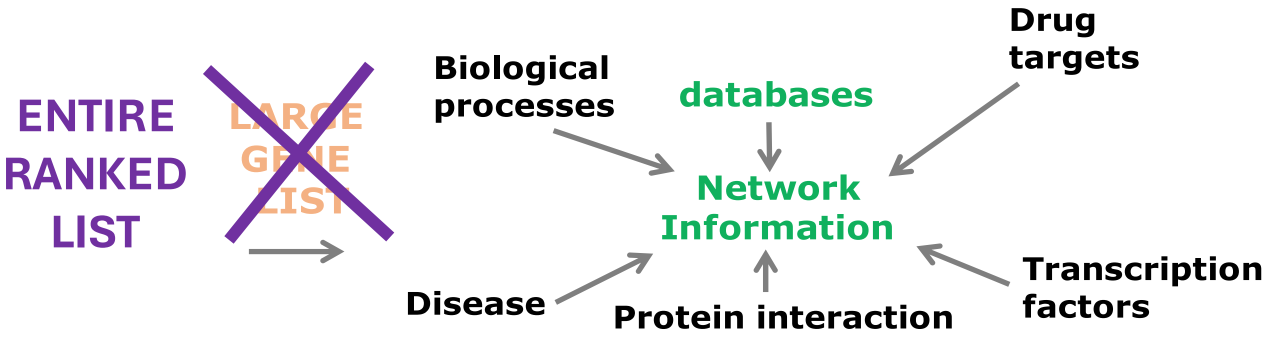
gene expression
methylation, CNVs

? ? ?

~~LARGE GENE LIST~~

ENTIRE
RANKED
LIST

Step2



ENTIRE
RANKED
LIST

~~LARGE GENE LIST~~

Biological processes

databases

Drug targets

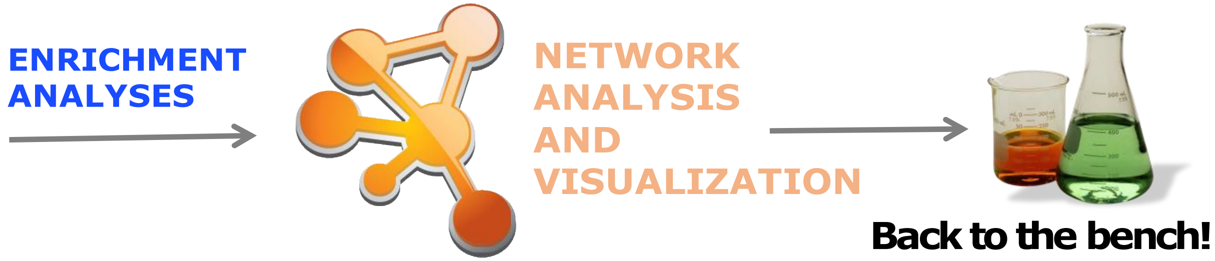
Network Information

Disease

Protein interaction

Transcription factors

Step3



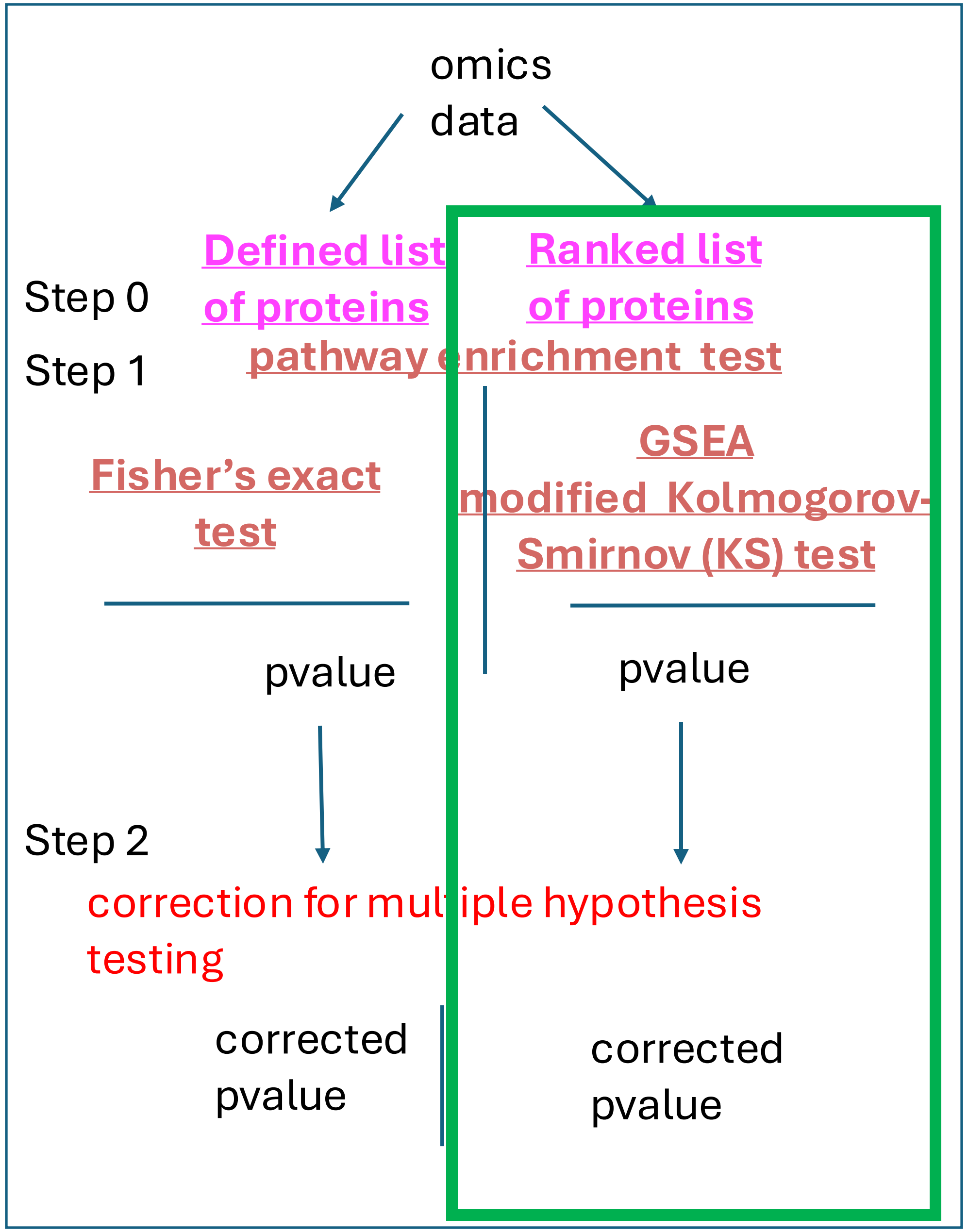
ENRICHMENT ANALYSES

NETWORK ANALYSIS AND VISUALIZATION

Back to the bench!

Two types of gene lists : defined or ranked list.

- Fisher's Exact Test, aka Hypergeometric Test
- GSEA for ranked lists.
- Multiple test corrections:
  - Bonferroni correction
  - False Discovery Rate computation using Benjamini-Hochberg procedure



Why test enrichment with ranked gene lists?

- Possible problems with thresholded gene list test include:
  - No “natural” value for the threshold. Avoids arbitrary cut-offs.
  - Different results at different threshold settings
  - Useful for Noisy or Weak Signals
    - Particularly valuable when no genes are strongly differentially expressed, but **pathways** still show directional change.
  - Possible loss of statistical power due to the thresholding
    - No resolution between significant signals with different strengths
    - Weak signals are ignored.

** reduces bias, increases reproducibility, and improves sensitivity for detecting real biological patterns

Gene Set Enrichment Analysis - GSEA

- Original paper published in Nature Genetics 2003 (Mootha el al) and modified and republished in PNAS 2005 (Subramanian et al)
- In the original paper, Mootha et al (2003) studied diabetes and identified that their gene list was significantly enriched in a pathway called “oxidative phosphorylation”.
- The particularity of this finding was that the individual genes in this pathway were only down-regulated by a small amount but the addition of all these subtle signals had a great impact on the pathway.
- They validated their finding experimentally.

GSEA score calculation

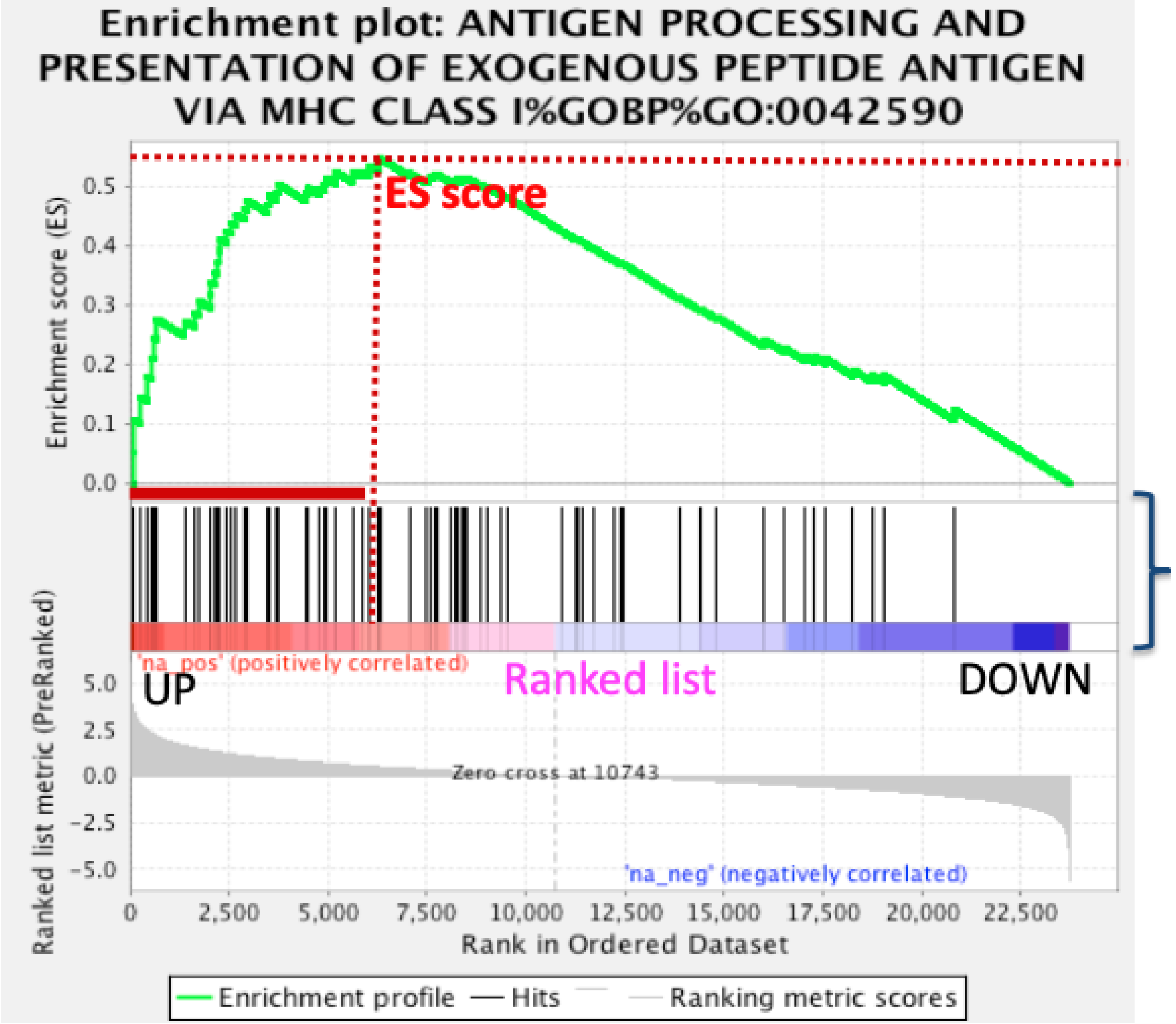
modified Kolmogorov Smirnov test (KS test)

Ranked gene list

UP

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	-15.6
TAP1	-19
PSMB9	-19.7

DOWN



1. Walk down the ranked list of all genes.

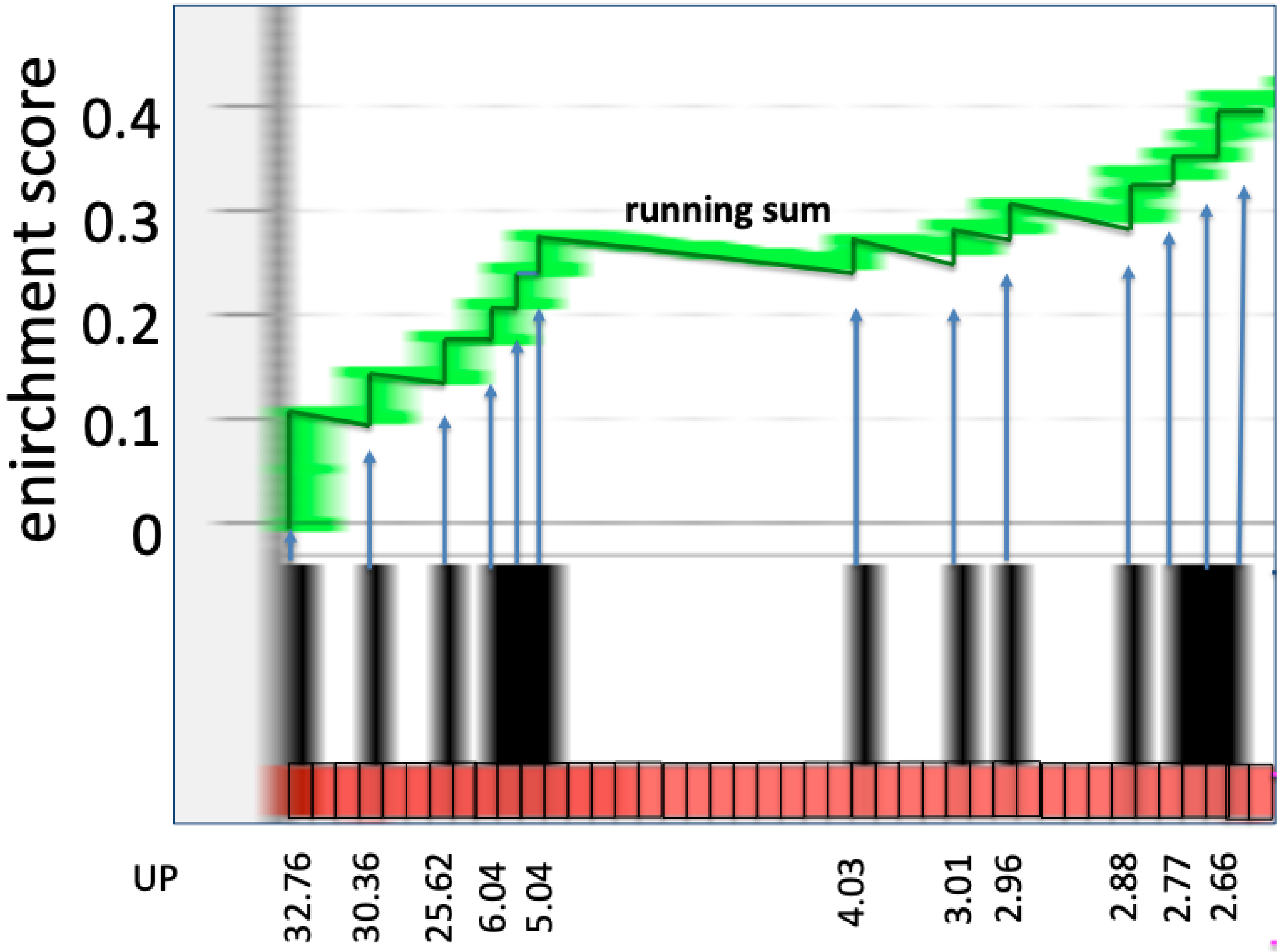
2. Increase a running-sum when a gene is in the gene set.

3. Decrease the running-sum when a gene is not.

4. The **maximum deviation from zero** of this running-sum is the **enrichment score (ES)**

5. Any gene contributing the ES before the maximum deviation is considered part of the **leading edge.**

Weighted - Kolmogorov Smirnov test (KS test)



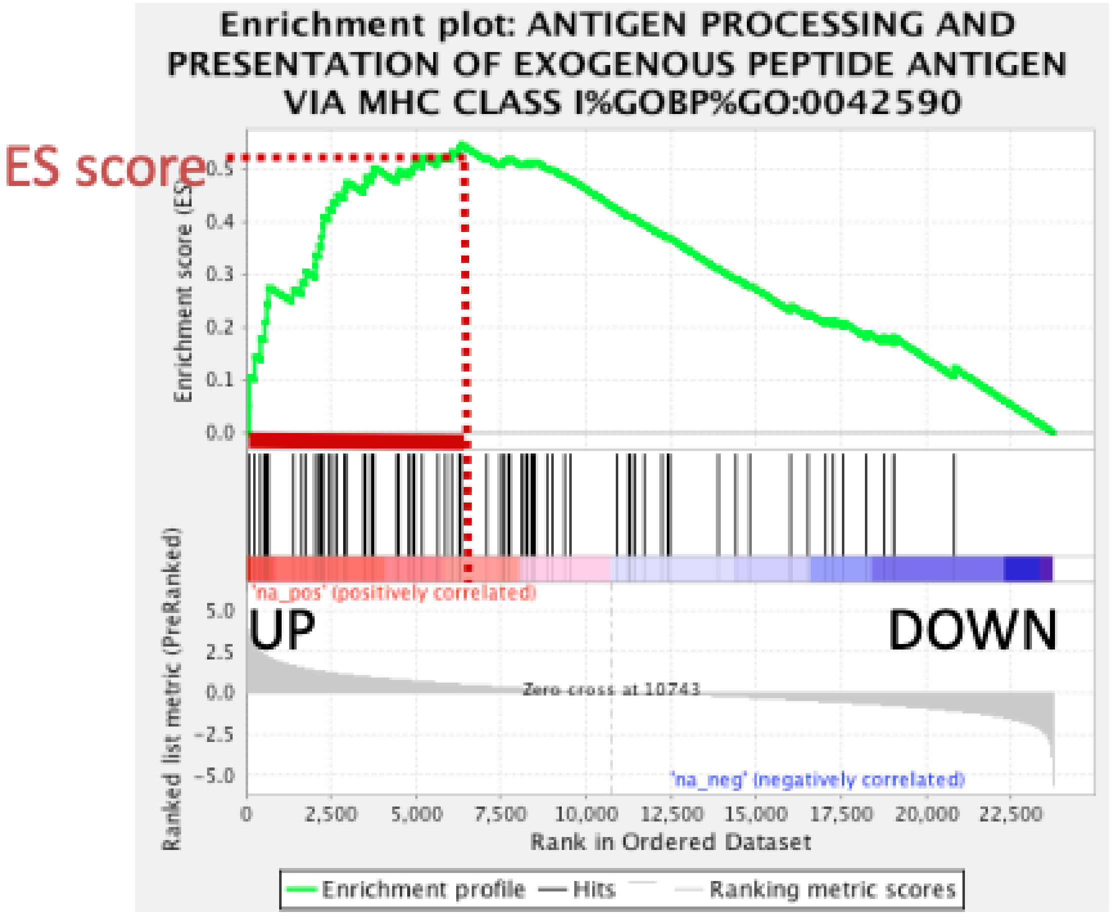
Ranks

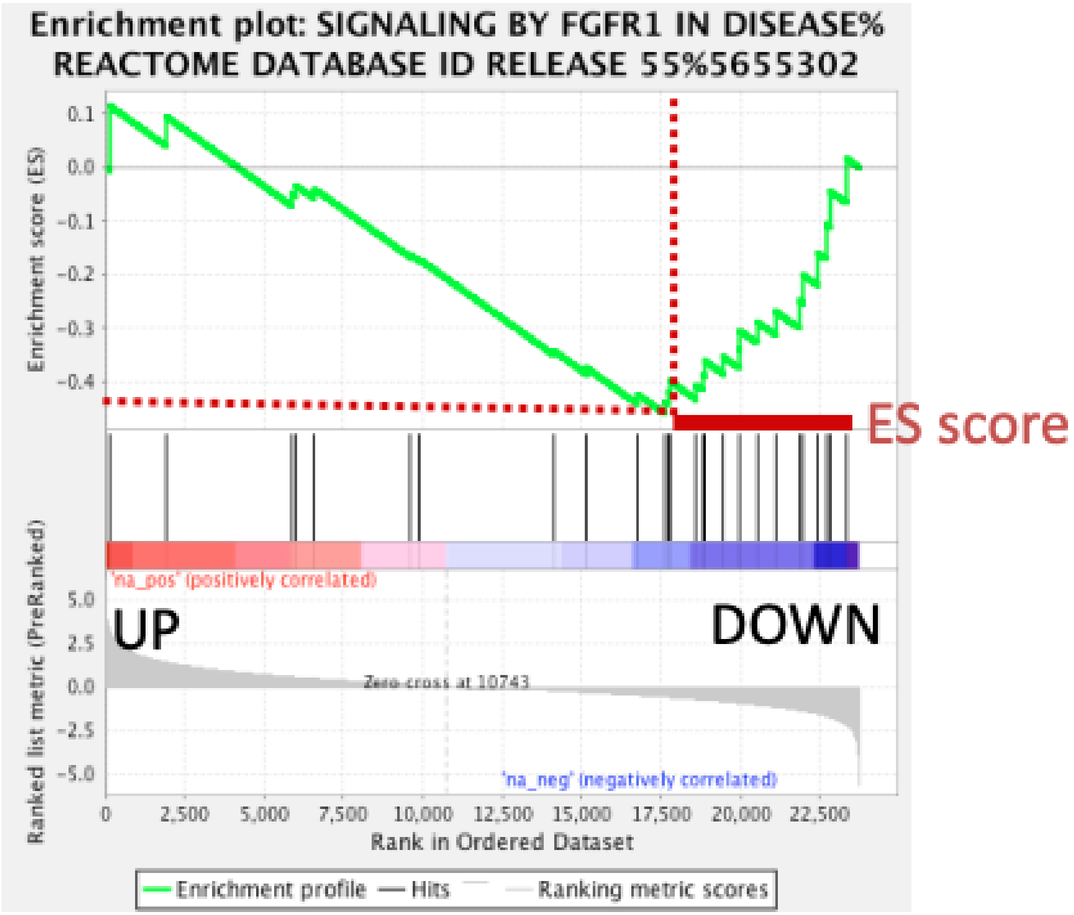
Weight
p=1 (default)

Options:
p=0 no weighting (all genes contribute equally, like Kolmogorov–Smirnov test)
p=1 linear weighting (default)
p>1 more emphasis on top-ranked genes

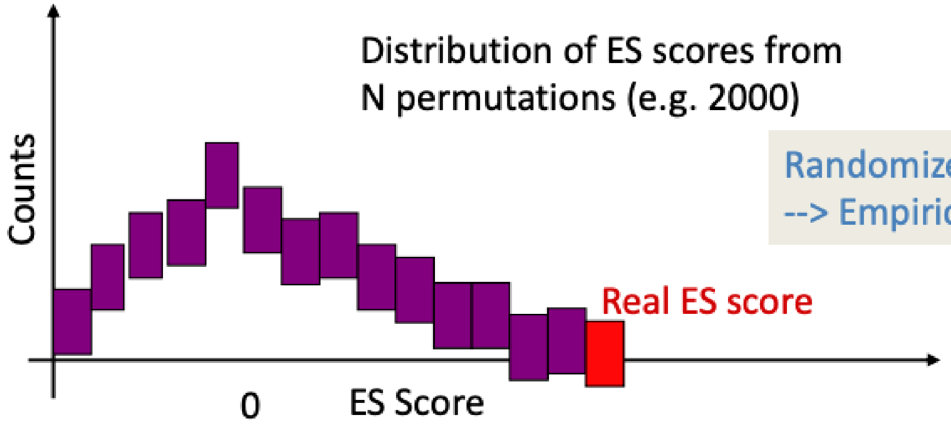
Found/not found in gene-set (pathway)

Positive and Negative enrichment scores





Going from ES, NES, p-value and FDR



Randomized with ES score ≥ real: 4 / 2,000
--> Empirical p-value = 0.002

- Randomize the ranked list and compute ES values for each pathway
- Repeat n times (by default 1000, but consider using higher)
- P-value = empirical p-value → (number of randomized ES scores > real value) / num of permutations
- NES = Normalized ES →

$$NES = \frac{ES}{\text{mean}(|ES_{null}|)}$$

- FDR = false discovery Rate is a multiple testing correction that estimates **how likely it is that a gene set with a given NES is a false positive.**

$$FDR(NES_{obs}) = \frac{\text{Number of null NES} \geq \text{NES}_{obs}}{\text{Number of observed NES} \geq \text{NES}_{obs}}$$

Which files do we need to run GSEA?

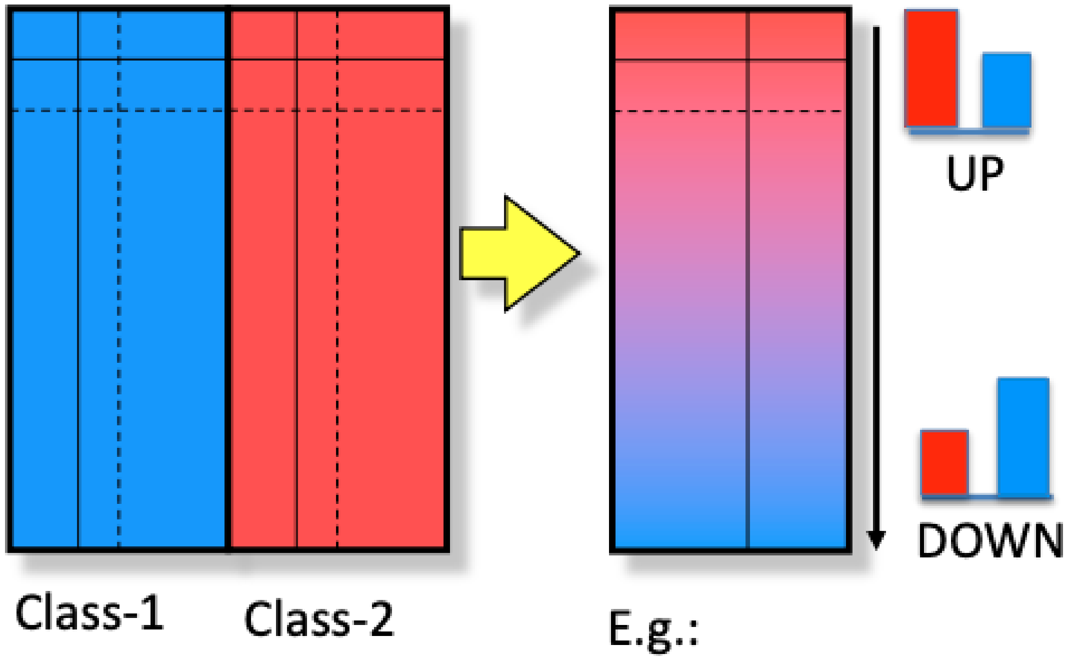
- A ranked list of genes called the rank file
  - this is a text file (tab separated) that should be renamed to end with the extension .rnk
  - This file has 2 columns :
    - gene identifier
    - ranking values
- A file called a .gmt file that contains the pathway data base (the gene-sets)
  - this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

What is a ranked gene list?

Two-class design: ranked gene list

Expression Matrix

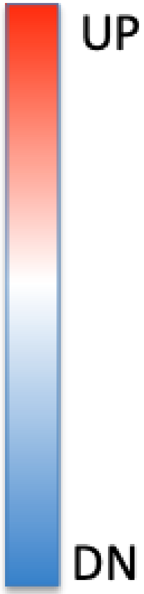
Genes Ranked by Differential Statistic



E.g.:
Fold change
Log (ratio)
t values from t-test

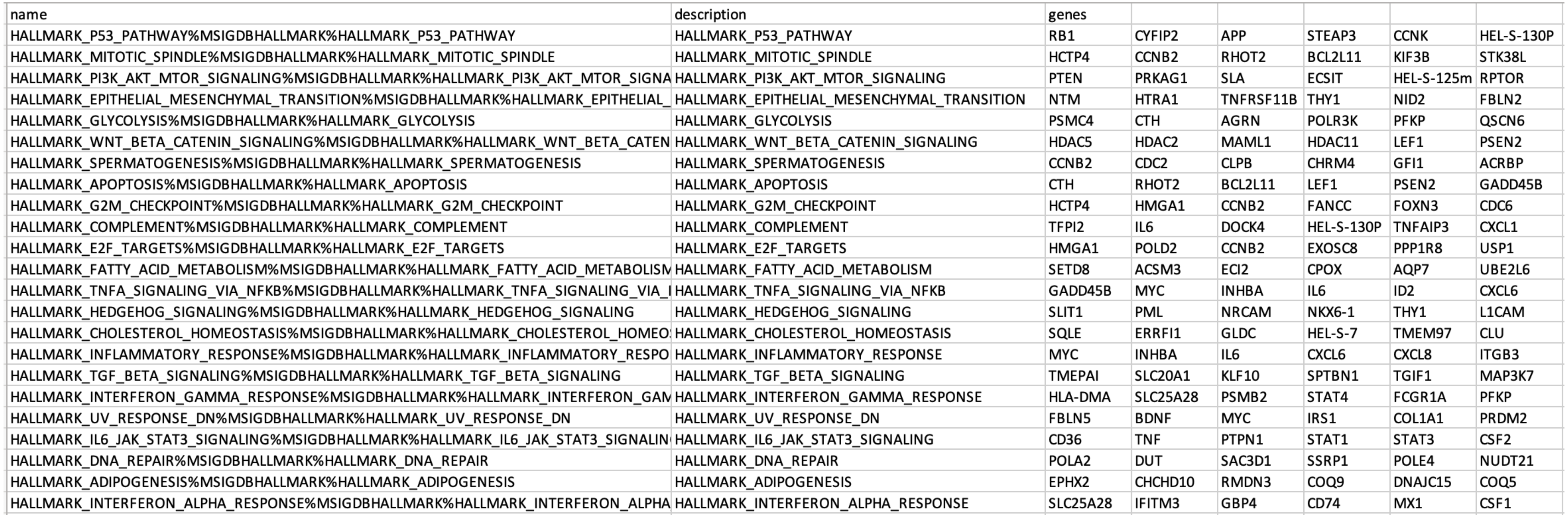
Ranking score = sign(logFC)*-log10(pvalue)

	LogFC	Pvalue	score
BGN	+1	1.73E-33	32.76
ANTXR1	+1	4.39E-31	30.36
FZD1	+1	4.41E-30	29.36
COL16A1	+1	1.33E-29	28.88
KLF3	+1	8.32E-02	1.08
RASEF	+1	9.01E-01	0.05
ISOC1	+1	9.01E-01	0.05
ANO1	+1	9.01E-01	0.04
CBWD3	-1	8.18E-02	-1.09
GBP4	-1	2.45E-16	-15.61
TAP1	-1	1.04E-19	-18.98
PSMB9	-1	1.84E-20	-19.73



What does a .gmt file look like?

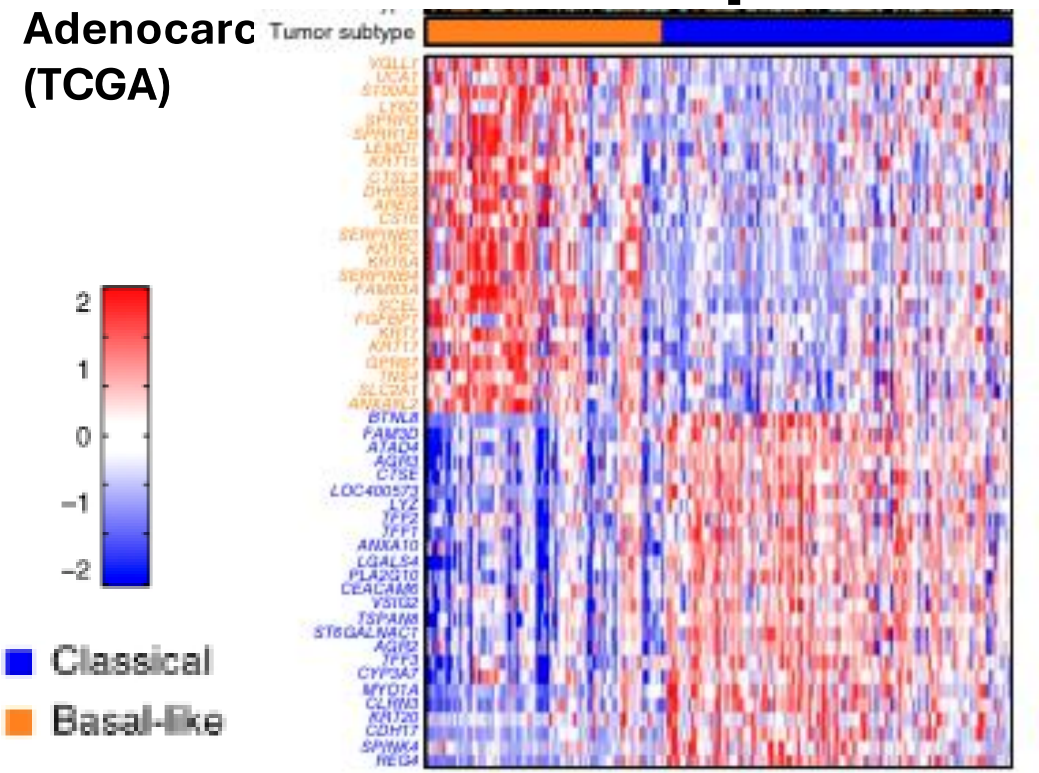
- Already introduced in part 1 but just for a refresher
  - It is a tab delimited text file with extension .gmt that contains pathways and their associated genes



name	description	genes					
HALLMARK_P53_PATHWAY%MSIGDBHALLMARK%HALLMARK_P53_PATHWAY	HALLMARK_P53_PATHWAY	RB1	CYFIP2	APP	STEAP3	CCNK	HEL-S-130P
HALLMARK_MITOTIC_SPINDLE%MSIGDBHALLMARK%HALLMARK_MITOTIC_SPINDLE	HALLMARK_MITOTIC_SPINDLE	HCTP4	CCNB2	RHOT2	BCL2L11	KIF3B	STK38L
HALLMARK_PI3K_AKT_MTOR_SIGNALING%MSIGDBHALLMARK%HALLMARK_PI3K_AKT_MTOR_SIGNA	HALLMARK_PI3K_AKT_MTOR_SIGNALING	PTEN	PRKAG1	SLA	ECSIT	HEL-S-125m	RPTOR
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION%MSIGDBHALLMARK%HALLMARK_EPITHELIAL_	HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	NTM	HTRA1	TNFRSF11B	THY1	NID2	FBLN2
HALLMARK_GLYCOLYSIS%MSIGDBHALLMARK%HALLMARK_GLYCOLYSIS	HALLMARK_GLYCOLYSIS	PSMC4	CTH	AGRN	POLR3K	PFKP	QSCN6
HALLMARK_WNT_BETA_CATENIN_SIGNALING%MSIGDBHALLMARK%HALLMARK_WNT_BETA_CATEN	HALLMARK_WNT_BETA_CATENIN_SIGNALING	HDAC5	HDAC2	MAML1	HDAC11	LEF1	PSEN2
HALLMARK_SPERMATOGENESIS%MSIGDBHALLMARK%HALLMARK_SPERMATOGENESIS	HALLMARK_SPERMATOGENESIS	CCNB2	CDC2	CLPB	CHRM4	GFI1	ACRBP
HALLMARK_APOPTOSIS%MSIGDBHALLMARK%HALLMARK_APOPTOSIS	HALLMARK_APOPTOSIS	CTH	RHOT2	BCL2L11	LEF1	PSEN2	GADD45B
HALLMARK_G2M_CHECKPOINT%MSIGDBHALLMARK%HALLMARK_G2M_CHECKPOINT	HALLMARK_G2M_CHECKPOINT	HCTP4	HMGA1	CCNB2	FANCC	FOXN3	CDC6
HALLMARK_COMPLEMENT%MSIGDBHALLMARK%HALLMARK_COMPLEMENT	HALLMARK_COMPLEMENT	TFPI2	IL6	DOCK4	HEL-S-130P	TNFAIP3	CXCL1
HALLMARK_E2F_TARGETS%MSIGDBHALLMARK%HALLMARK_E2F_TARGETS	HALLMARK_E2F_TARGETS	HMGA1	POLD2	CCNB2	EXOSC8	PPP1R8	USP1
HALLMARK_FATTY_ACID_METABOLISM%MSIGDBHALLMARK%HALLMARK_FATTY_ACID_METABOLISM	HALLMARK_FATTY_ACID_METABOLISM	SETD8	ACSM3	ECI2	CPOX	AQP7	UBE2L6
HALLMARK_TNFA_SIGNALING_VIA_NFKB%MSIGDBHALLMARK%HALLMARK_TNFA_SIGNALING_VIA_I	HALLMARK_TNFA_SIGNALING_VIA_NFKB	GADD45B	MYC	INHBA	IL6	ID2	CXCL6
HALLMARK_HEDGEHOG_SIGNALING%MSIGDBHALLMARK%HALLMARK_HEDGEHOG_SIGNALING	HALLMARK_HEDGEHOG_SIGNALING	SLIT1	PML	NRCAM	NKX6-1	THY1	L1CAM
HALLMARK_CHOLESTEROL_HOMEOSTASIS%MSIGDBHALLMARK%HALLMARK_CHOLESTEROL_HOMEO	HALLMARK_CHOLESTEROL_HOMEOSTASIS	SQLE	ERRFI1	GLDC	HEL-S-7	TMEM97	CLU
HALLMARK_INFLAMMATORY_RESPONSE%MSIGDBHALLMARK%HALLMARK_INFLAMMATORY_RESPO	HALLMARK_INFLAMMATORY_RESPONSE	MYC	INHBA	IL6	CXCL6	CXCL8	ITGB3
HALLMARK_TGF_BETA_SIGNALING%MSIGDBHALLMARK%HALLMARK_TGF_BETA_SIGNALING	HALLMARK_TGF_BETA_SIGNALING	TMEPAI	SLC20A1	KLF10	SPTBN1	TGIF1	MAP3K7
HALLMARK_INTERFERON_GAMMA_RESPONSE%MSIGDBHALLMARK%HALLMARK_INTERFERON_GAM	HALLMARK_INTERFERON_GAMMA_RESPONSE	HLA-DMA	SLC25A28	PSMB2	STAT4	FCGR1A	PFKP
HALLMARK_UV_RESPONSE_DN%MSIGDBHALLMARK%HALLMARK_UV_RESPONSE_DN	HALLMARK_UV_RESPONSE_DN	FBLN5	BDNF	MYC	IRS1	COL1A1	PRDM2
HALLMARK_IL6_JAK_STAT3_SIGNALING%MSIGDBHALLMARK%HALLMARK_IL6_JAK_STAT3_SIGNALIN	HALLMARK_IL6_JAK_STAT3_SIGNALING	CD36	TNF	PTPN1	STAT1	STAT3	CSF2
HALLMARK_DNA_REPAIR%MSIGDBHALLMARK%HALLMARK_DNA_REPAIR	HALLMARK_DNA_REPAIR	POLA2	DUT	SAC3D1	SSRP1	POLE4	NUDT21
HALLMARK_ADIPOGENESIS%MSIGDBHALLMARK%HALLMARK_ADIPOGENESIS	HALLMARK_ADIPOGENESIS	EPHX2	CHCHD10	RMDN3	COQ9	DNAJC15	COQ5
HALLMARK_INTERFERON_ALPHA_RESPONSE%MSIGDBHALLMARK%HALLMARK_INTERFERON_ALPHA	HALLMARK_INTERFERON_ALPHA_RESPONSE	SLC25A28	IFITM3	GBP4	CD74	MX1	CSF1

EXAMPLE WITH A RANKED GENE LIST

Dataset
Pancreatic
Ductal
Adenocarc
(TCGA)



Moffitt, R., Marayati, R., Flate, E. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* **47**, 1168–1178 (2015)

Basal **vs** **Classical**

↓

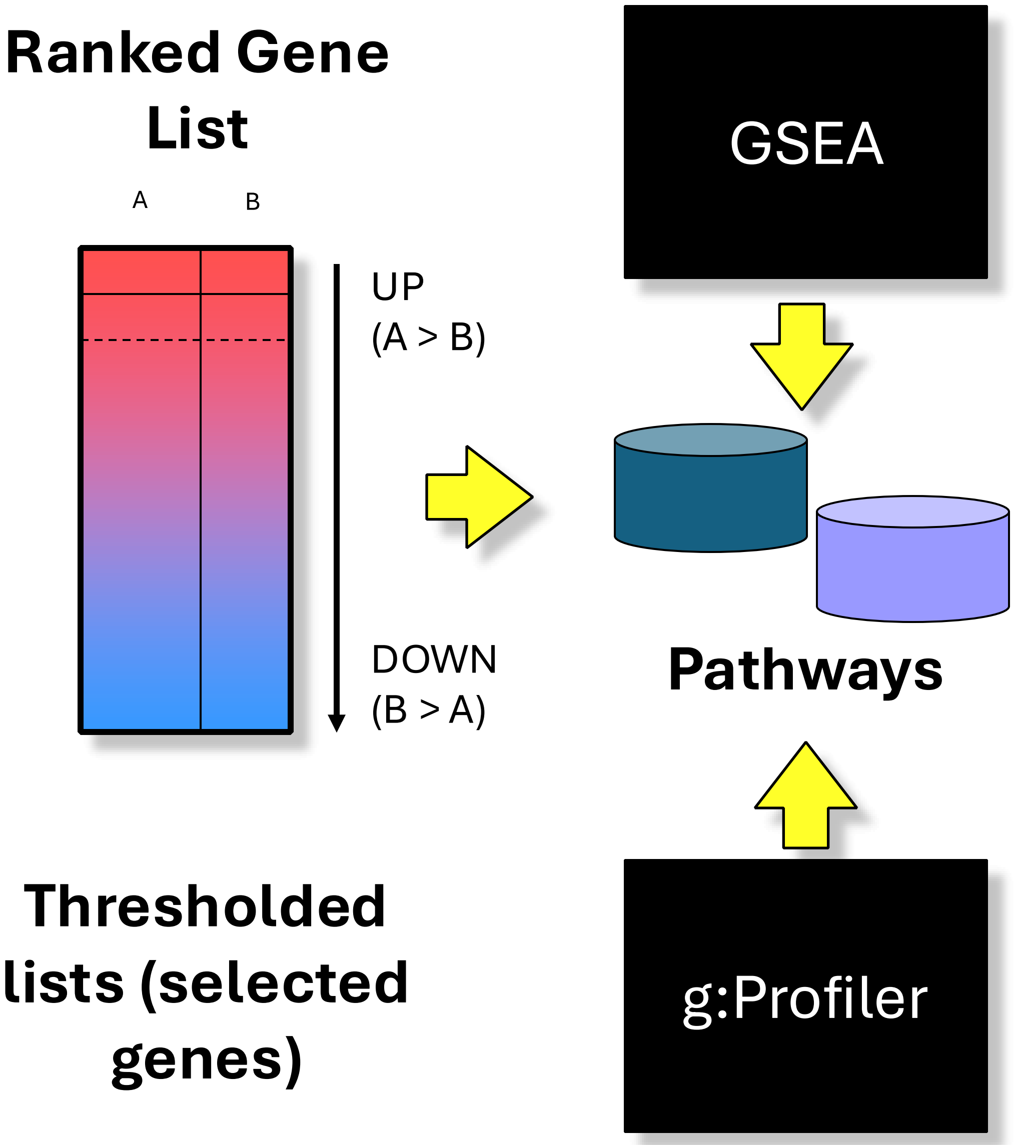
Differential expression (edgeR)

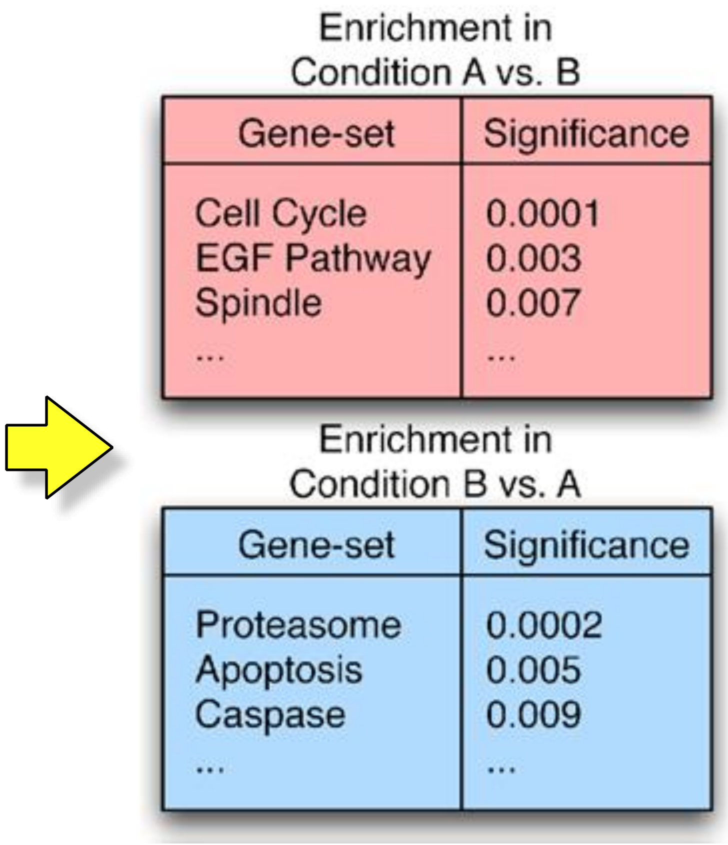
Rank file

↓

Gene-Set Enrichment Analysis (fGSEA)

Pathway Enrichment Analysis





Enrichment in Condition A vs. B

Gene-set	Significance
Cell Cycle	0.0001
EGF Pathway	0.003
Spindle	0.007
...	...

Enrichment in Condition B vs. A

Gene-set	Significance
Proteasome	0.0002
Apoptosis	0.005
Caspase	0.009
...	...

Merico D, Isserlin R, Stueker O, Emili A, Bader GD Enrichment map: a network-based method for gene-set enrichment visualization and interpretation PLoS One. 2010 Nov 15;5(11):e13984

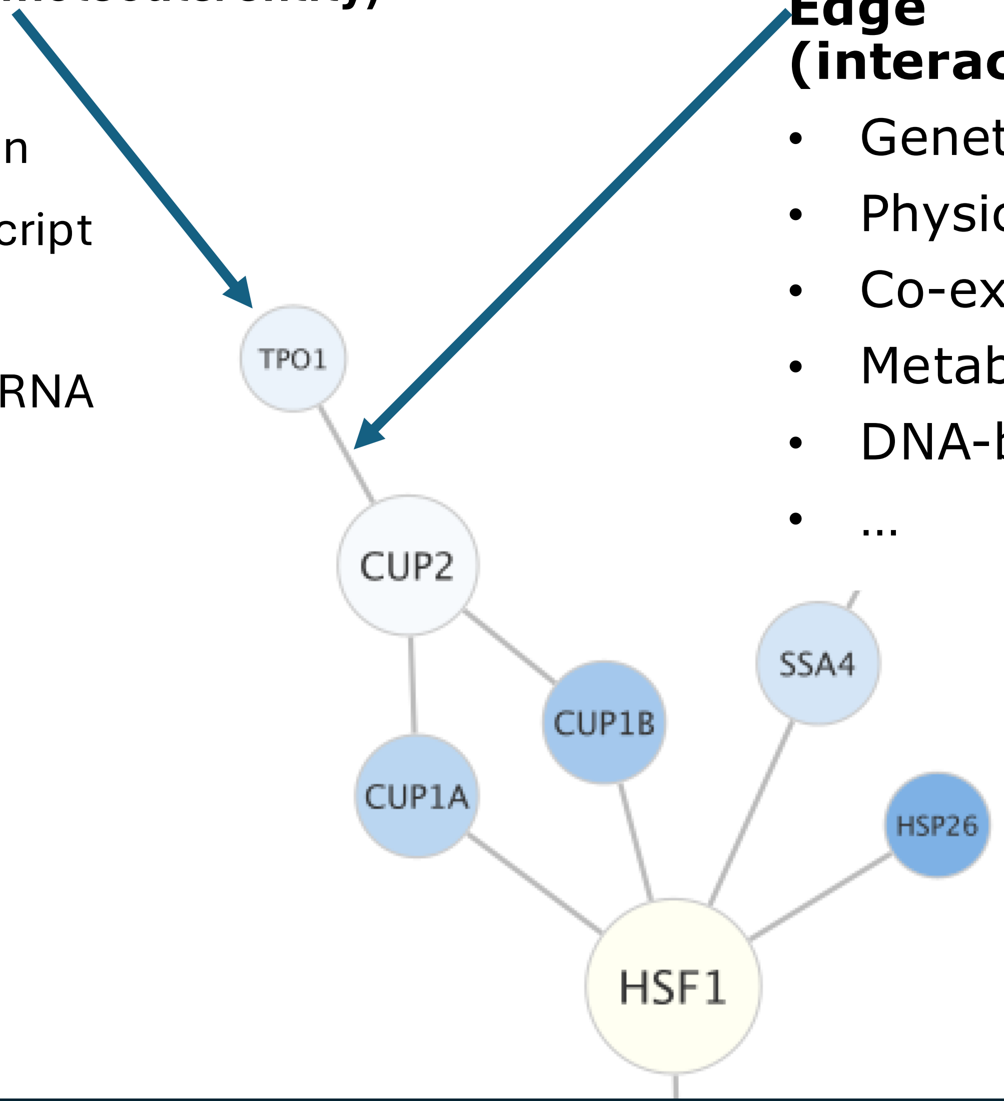
Network Basics

Node (molecule/entity)

- Gene
- Protein
- Transcript
- Drug
- MicroRNA
- …

Edge (interaction/relationship)

- Genetic interaction
- Physical protein interaction
- Co-expression
- Metabolic reaction
- DNA-binding
- …



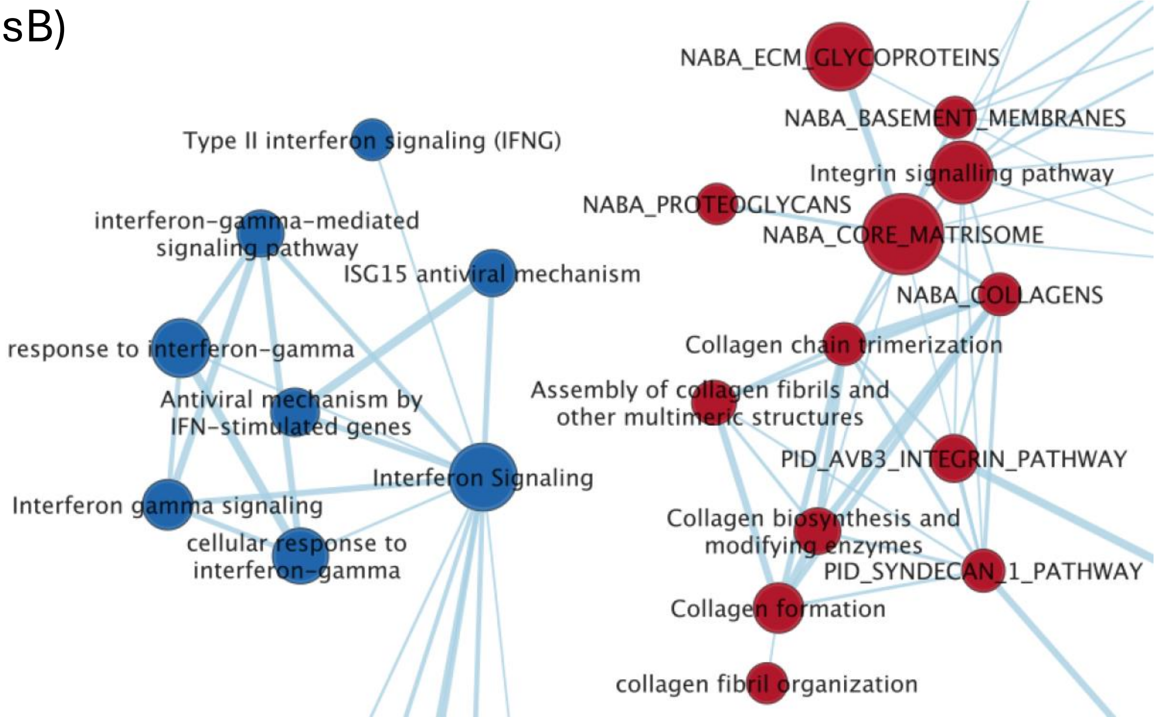
Enrichment Map Basics

Node (molecule/entity)

- Pathway or geneset
- Size is correlated to number of genes in set
- Color indicates class in below example (for example Up/Down, classA/classB)

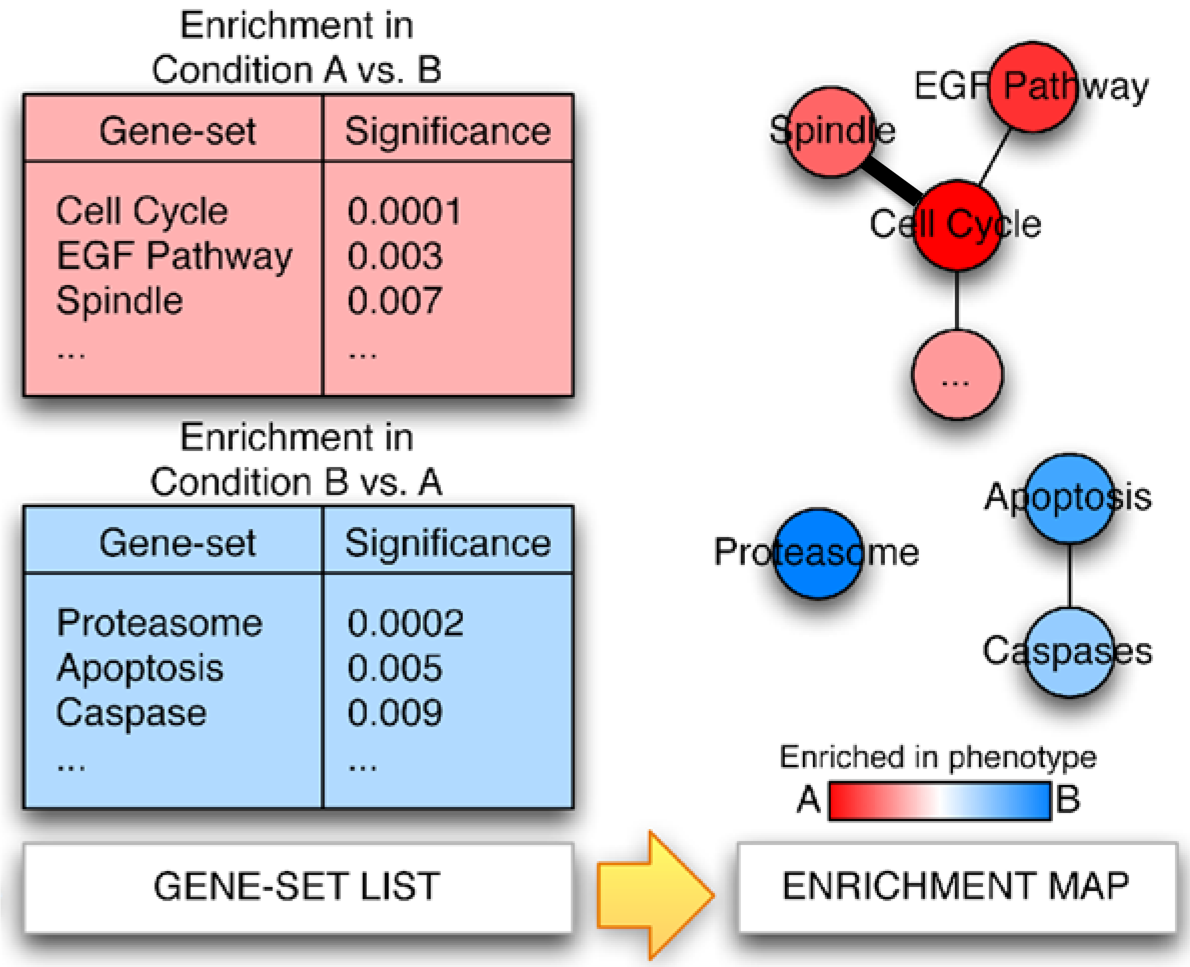
Edge (interaction/relationship)

- Degree of overlap between two genesets
- The more genes two pathways have in common the thicker/stronger the connection



Type II interferon signaling (IFNG)
interferon-gamma-mediated signaling pathway
ISG15 antiviral mechanism
response to interferon-gamma
Antiviral mechanism by IFN-stimulated genes
Interferon Signaling
Interferon gamma signaling
cellular response to interferon-gamma
NABA_ECM_GLYCOPROTEINS
NABA_BASEMENT_MEMBRANES
Integrin signalling pathway
NABA_PROTEOGLYCANS
NABA_CORE_MATRISOME
NABA_COLLAGENS
Collagen chain trimerization
Assembly of collagen fibrils and other multimeric structures
PID_AVB3_INTEGRIN_PATHWAY
Collagen biosynthesis and modifying enzymes
PID_SYNDECAN_1_PATHWAY
Collagen formation
collagen fibril organization

Enrichment Map



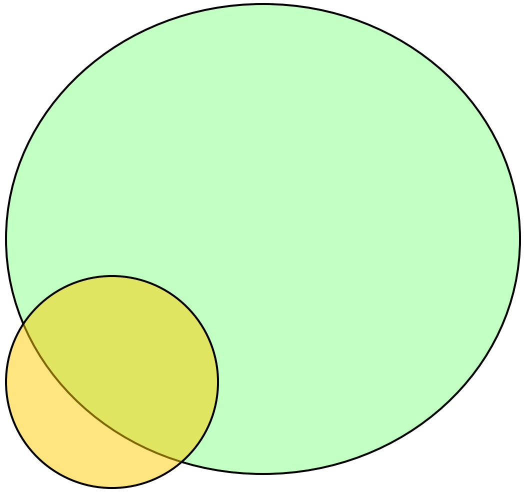
Enrichment in Condition A vs. B

Gene-set	Significance
Cell Cycle EGF Pathway Spindle ...	0.0001 0.003 0.007 ...

Enrichment in Condition B vs. A

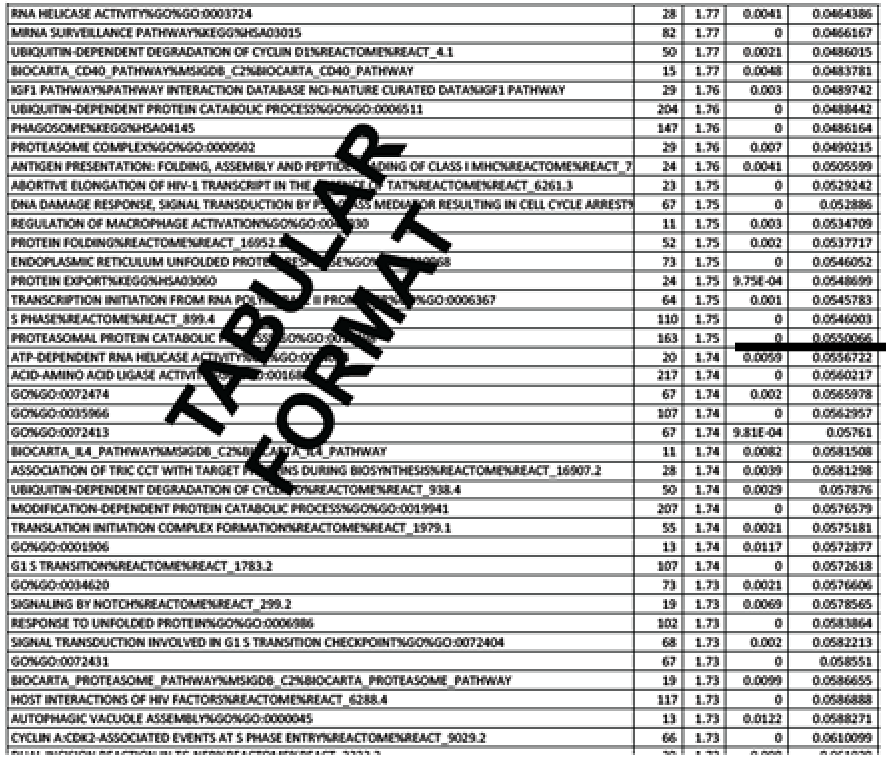
Gene-set	Significance
Proteasome Apoptosis Caspase ...	0.0002 0.005 0.009 ...

Overlap



$$\frac{|A \cap B|}{\min(|A|,|B|)}$$

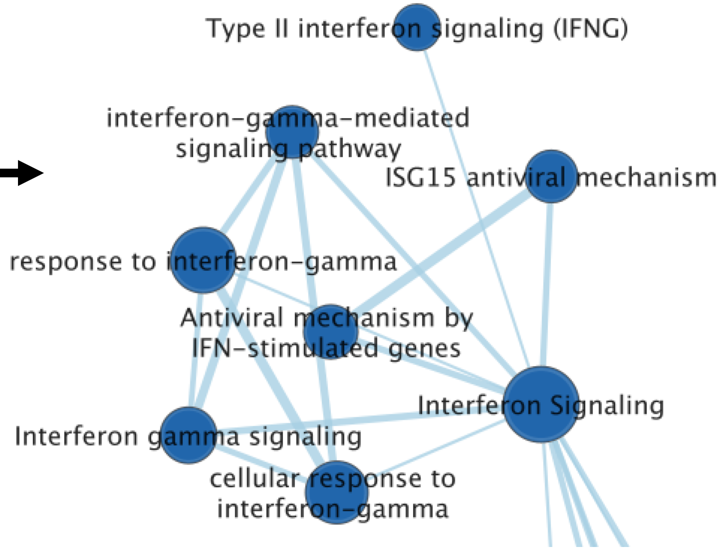
Typical Output



TABULAR FORMAT

RNA HELICASE ACTIVITY%GO%GO:0003724	28	1.77	0.0041	0.0464386
MRNA SURVEILLANCE PATHWAY%KEGG%HSA03015	82	1.77	0	0.0466167
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN D1%REACTOME%REACT_4.1	50	1.77	0.0021	0.0486015
BIOCARTA_CD40_PATHWAY%MSIGDB_C2%BIOCARTA_CD40_PATHWAY	15	1.77	0.0048	0.0483781
IGF1 PATHWAY%PATHWAY INTERACTION DATABASE NCI-NATURE CURATED DATA%IGF1 PATHWAY	29	1.76	0.003	0.0489742
UBIQUITIN-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0006511	204	1.76	0	0.0488442
PHAGOSOME%KEGG%HSA04145	147	1.76	0	0.0486164
PROTEASOME COMPLEX%GO%GO:0000502	29	1.76	0.007	0.0490215
ANTIGEN PRESENTATION: FOLDING, ASSEMBLY AND PEPTID[illegible]DING OF CLASS I MHC%REACTOME%REACT_7	24	1.76	0.0041	0.0505599
ABORTIVE ELONGATION OF HIV-1 TRANSCRIPT IN THE [illegible] TAT%REACTOME%REACT_6261.3	23	1.75	0	0.0529242
DNA DAMAGE RESPONSE, SIGNAL TRANSDUCTION BY P[illegible]SS MEDIA[illegible]OR RESULTING IN CELL CYCLE ARREST%	67	1.75	0	0.052886
REGULATION OF MACROPHAGE ACTIVATION%GO%GO:00[illegible]30	11	1.75	0.003	0.0534709
PROTEIN FOLDING%REACTOME%REACT_16952.[illegible]	52	1.75	0.002	0.0537717
ENDOPLASMIC RETICULUM UNFOLDED PROTE[illegible]	73	1.75	0	0.0546052
PROTEIN EXPORT%KEGG%HSA03060	24	1.75	9.75E-04	0.0548699
TRANSCRIPTION INITIATION FROM RNA POL[illegible] II PRO[illegible]%GO:0006367	64	1.75	0.001	0.0545783
S PHASE%REACTOME%REACT_899.4	110	1.75	0	0.0546003
PROTEASOMAL PROTEIN CATABOLIC [illegible]%GO%GO:0[illegible]	163	1.75	0	0.0550066
ATP-DEPENDENT RNA HELICASE ACTIVITY[illegible]GO:00[illegible]	20	1.74	0.0059	0.0556722
ACID-AMINO ACID LIGASE ACTIVIT[illegible]:0016[illegible]	217	1.74	0	0.0560217
GO%GO:0072474	67	1.74	0.002	0.0565978
GO%GO:0035966	107	1.74	0	0.0562957
GO%GO:0072413	67	1.74	9.81E-04	0.05761
BIOCARTA_IL4_PATHWAY%MSIGDB_C2%BI[illegible]_PATHWAY	11	1.74	0.0082	0.0581508
ASSOCIATION OF TRIC CCT WITH TARGET P[illegible]NS DURING BIOSYNTHESIS%REACTOME%REACT_16907.2	28	1.74	0.0039	0.0581298
UBIQUITIN-DEPENDENT DEGRADATION OF CYC[illegible]%REACTOME%REACT_938.4	50	1.74	0.0029	0.057876
MODIFICATION-DEPENDENT PROTEIN CATABOLIC PROCESS%GO%GO:0019941	207	1.74	0	0.0576579
TRANSLATION INITIATION COMPLEX FORMATION%REACTOME%REACT_1979.1	55	1.74	0.0021	0.0575181
GO%GO:0001906	13	1.74	0.0117	0.0572877
G1 S TRANSITION%REACTOME%REACT_1783.2	107	1.74	0	0.0572618
GO%GO:0034620	73	1.73	0.0021	0.0576606
SIGNALING BY NOTCH%REACTOME%REACT_299.2	19	1.73	0.0069	0.0578565
RESPONSE TO UNFOLDED PROTEIN%GO%GO:0006986	102	1.73	0	0.0583864
SIGNAL TRANSDUCTION INVOLVED IN G1 S TRANSITION CHECKPOINT%GO%GO:0072404	68	1.73	0.002	0.0582213
GO%GO:0072431	67	1.73	0	0.058551
BIOCARTA_PROTEASOME_PATHWAY%MSIGDB_C2%BIOCARTA_PROTEASOME_PATHWAY	19	1.73	0.0099	0.0586655
HOST INTERACTIONS OF HIV FACTORS%REACTOME%REACT_6288.4	117	1.73	0	0.0586888
AUTOPHAGIC VACUOLE ASSEMBLY%GO%GO:0000045	13	1.73	0.0122	0.0588271
CYCLIN A:CDK2-ASSOCIATED EVENTS AT S PHASE ENTRY%REACTOME%REACT_9029.2	66	1.73	0	0.0610099

Network Visualization



Type II interferon signaling (IFNG)
interferon−gamma−mediated signaling pathway
ISG15 antiviral mechanism
response to interferon−gamma
Antiviral mechanism by IFN−stimulated genes
Interferon Signaling
Interferon gamma signaling
cellular response to interferon−gamma

Each row is a gene-set (pathway).
It displays:

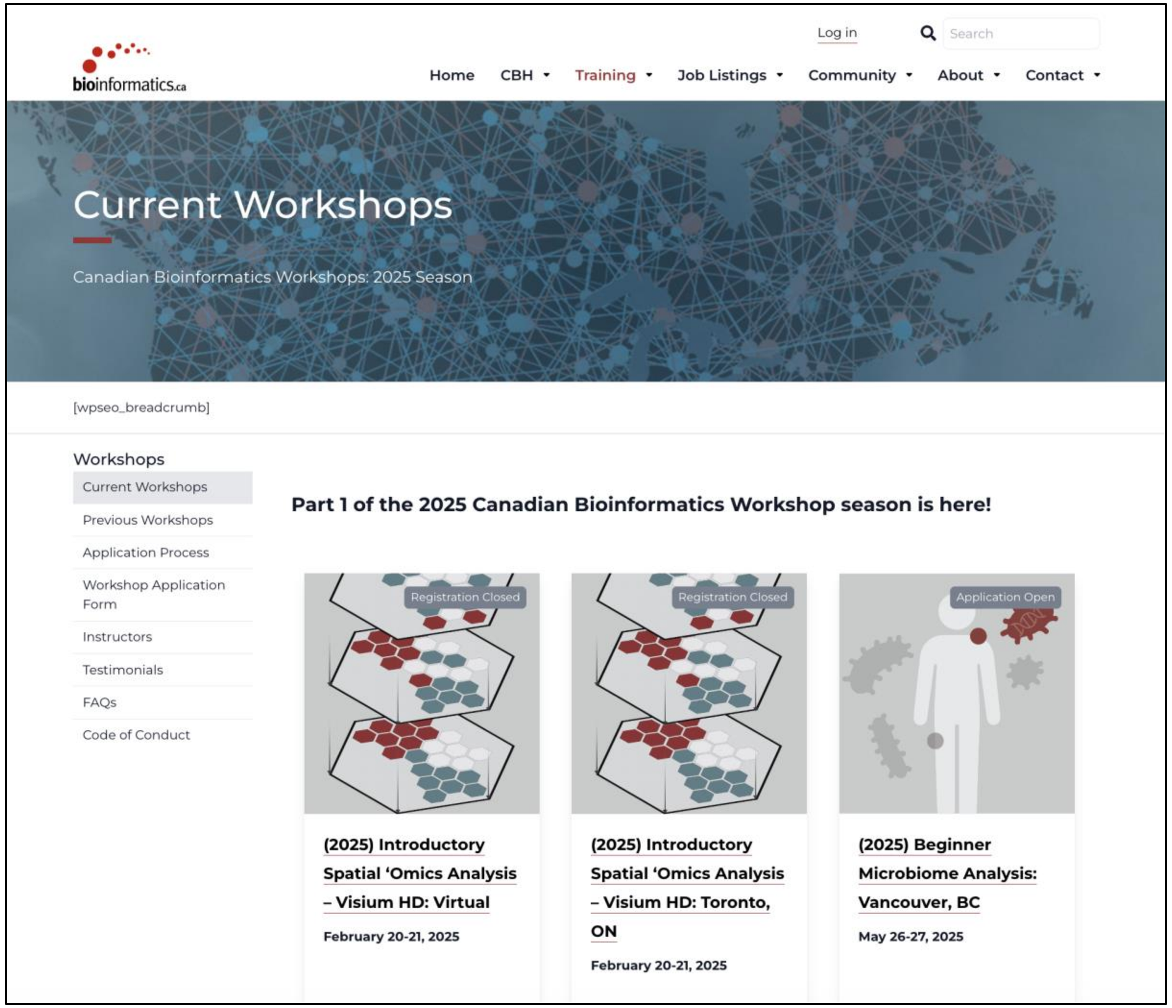
- a score associated with the magnitude of overlap between gene-set and gene list.
- a pvalue that estimates the significance of the enrichment (by chance or not).
- a corrected pvalue (FDR) that corrects for multiple hypothesis testing.

During the practical lab:

- We will run fGSEA from R
- Examine and Visualize the results in R
- Create output files that we can then use in Cytoscape

This module is taken from
CBW Pathway and Network Analysis Workshop 2024

2025 Canadian Bioinformatics Workshop
Please visit the website to see workshops presently offered



Current Workshops

Canadian Bioinformatics Workshops: 2025 Season

[wpseo_breadcrumb]

Workshops
- Current Workshops
- Previous Workshops
- Application Process
- Workshop Application Form
- Instructors
- Testimonials
- FAQs
- Code of Conduct

Part 1 of the 2025 Canadian Bioinformatics Workshop season is here!

(2025) Introductory Spatial 'Omics Analysis – Visium HD: Virtual
February 20-21, 2025

(2025) Introductory Spatial 'Omics Analysis – Visium HD: Toronto, ON
February 20-21, 2025

(2025) Beginner Microbiome Analysis: Vancouver, BC
May 26-27, 2025

https://bioinformatics.ca/workshops/current-workshops/

bioinformatics.ca

Time for practical lab

- https://baderlab.github.io/ComputeOntario_Pathways_2025/gsea-lab.html