Differential expression and downstream analysis
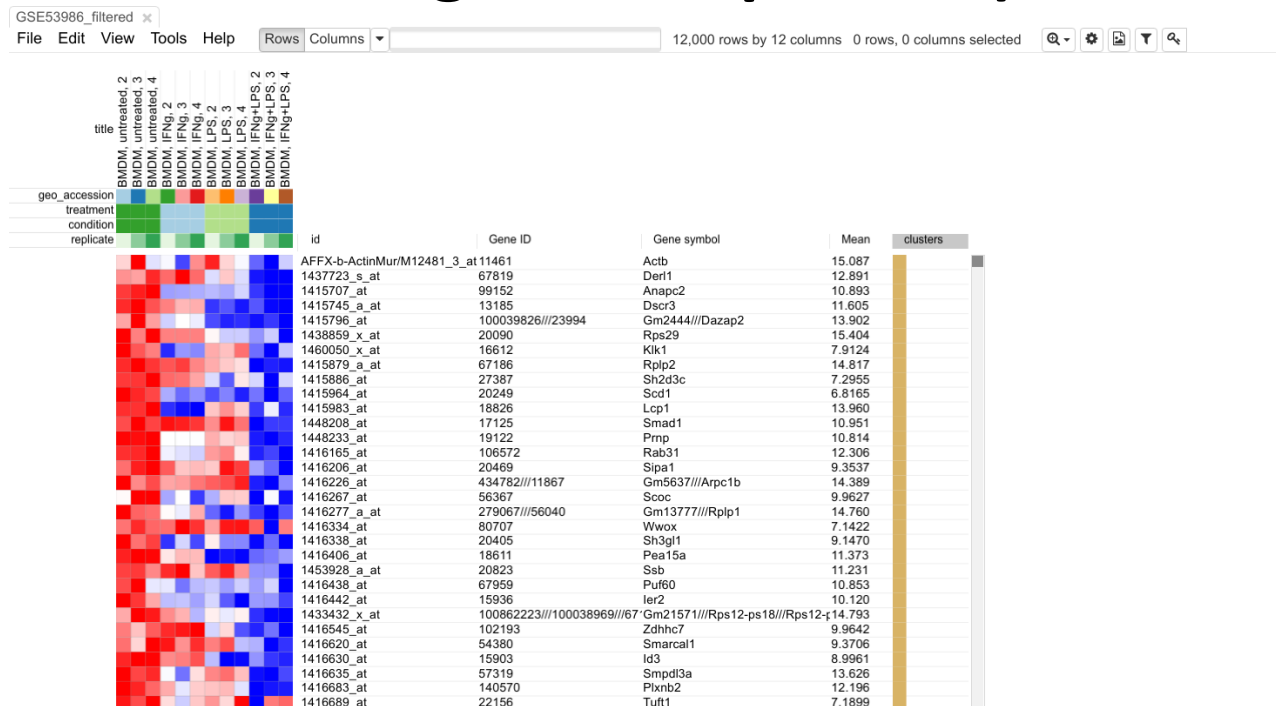
Alexey Sergushichev

2021-08-26, Tomsk

Outline

✅ Exploring gene expression datasets

✅ **Simple analysis methods**

✅ Working with public datasets

Open the filtered dataset from gct (or continue working in the open tab)

Differential expression

✔ Tools/Differential expression/limma

Field <-
treatment

untreated vs LPS

Differential expression: results

✅ Sort decreasing by "t"

✅ Save the results as GSE53986_Ctrl_vs_LPS.gct

Click to sort

P-value

✔ P-value - the probability that, when the **null hypothesis** is true, the statistical summary (such as the sample mean difference between two groups) would be equal to, or **more extreme** than, the actual observed results.
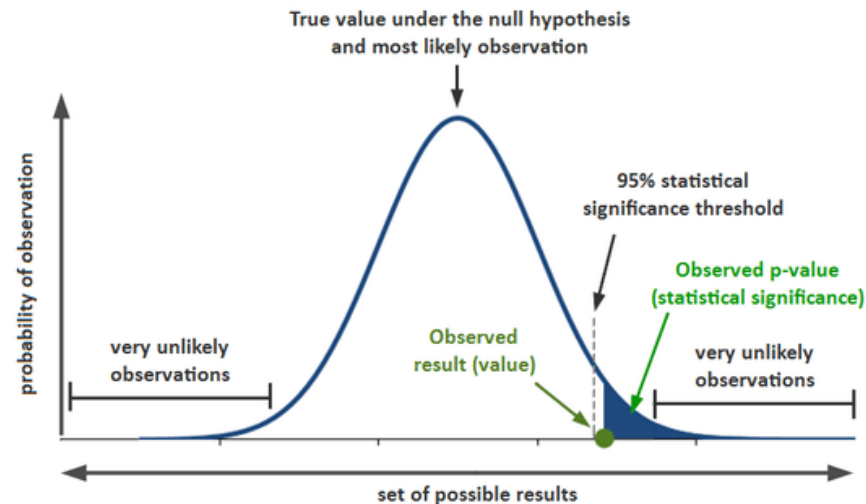
Differential expression p-value

✔ Null hypothesis: average gene expression is distributed in two conditions **and** the distributions follows a certain model

✔ Summary: t-statistic like

✔ When p-value is low, then we **reject** null hypothesis and assume genes are differentially expressed

Multiple hypothesis testing

✅ P-values are uniformly distributed when null hypothesis is true

✅ If there are no real differential expression at all, for 10^4 genes there will be P-values as low as 10^{-4}

✅ Using Benjamini-Hochberg adjustment procedure which limits False Discovery Rate (FDR)

https://xkcd.com/882/

Differential expression: results

✅ Sort decreasing by "t"

✅ Save the results as GSE53986_Ctrl_vs_LPS.gct

Click to sort

Sharing the results

- ✅ File/Get link to a dataset
- ✅ Open the link in another browser tab
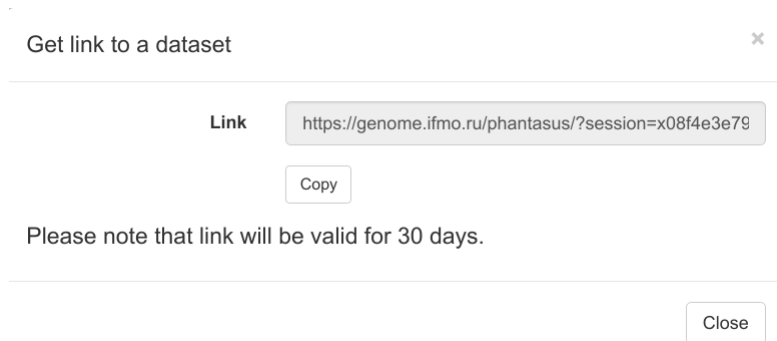
Get link to a dataset   ×

Link   https://genome.ifmo.ru/phantasus/?session=x08f4e3e79

Copy

Please note that link will be valid for 30 days.

Close

Pathway enrichment: using MSigDB tool

✅ Go to http://software.broadinstitute.org/gsea/msigdb/annotate.jsp

Put alsergbox@gmail.com here for now
Register later, it's free!

Pathway enrichment: selecting signature genes
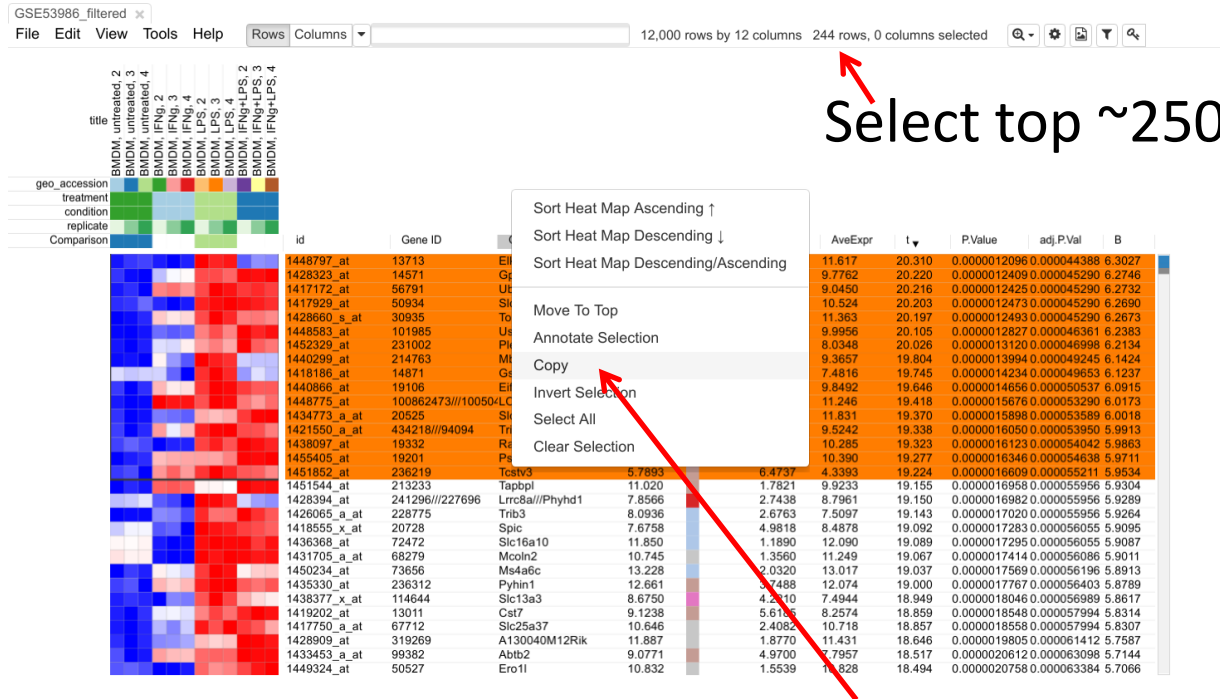
Select top ~250 genes

Right click on "Gene symbol" column
and click "Copy" (not Ctrl-C)

Pathway enrichment: using MSigDB tool

Paste genes here

Check "hallmark gene sets" and "Canonical pathways"

Click "compute overlaps"

Pathway enrichment results

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ❓	FDR q-value ❓
HALLMARK_INTERFERON_GAMMA_RESPONSE [200]	Genes up-regulated in response to IFNG [GeneID=3458].	67		$5.51\ e^{-112}$	$7.6\ e^{-109}$
HALLMARK_INTERFERON_ALPHA_RESPONSE [97]	Genes up-regulated in response to alpha interferon proteins.	34		$6.43\ e^{-57}$	$4.43\ e^{-54}$
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	49		$2.51\ e^{-39}$	$1.15\ e^{-36}$
HALLMARK_INFLAMMATORY_RESPONSE [200]	Genes defining inflammatory response.	30		$7.61\ e^{-38}$	$2.63\ e^{-35}$
HALLMARK_TNFA_SIGNALING_VIA_NFKB [200]	Genes regulated by NF-kB in response to TNF [GeneID=7124].	29		$3.76\ e^{-36}$	$1.04\ e^{-33}$
REACTOME_INTERFERON_SIGNALING [159]	Genes involved in Interferon Signaling	26		$6.39\ e^{-34}$	$1.47\ e^{-31}$
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	30		$9.04\ e^{-34}$	$1.78\ e^{-31}$
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING [64]	Genes involved in Interferon alpha/beta signaling	17		$1.85\ e^{-26}$	$3.19\ e^{-24}$
HALLMARK_ALLOGRAFT_REJECTION [200]	Genes up-regulated during transplant rejection.	18		$5.83\ e^{-19}$	$8.78\ e^{-17}$
REACTOME_INTERFERON_GAMMA_SIGNALING [63]	Genes involved in Interferon gamma signaling	13		$6.37\ e^{-19}$	$8.78\ e^{-17}$

gene/pathway matrix

Gene/geneset overlap matrix

Entrez Gene Id	Gene Symbol	HALLMARK_INTERFERON_GAMMA_RESPONSE	HALLMARK_INTERFERON_ALPHA_RESPONSE	REACTOME_IMMUNE_SYSTEM	HALLMARK_INFLAMMATORY_RESPONSE	HALLMARK_TNFA_SIGNALING_VIA_NFKB	REACTOME_INTERFERON_SIGNALING	REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	HALLMARK_ALLOGRAFT_REJECTION	REACTOME_INTERFERON_GAMMA_SIGNALING	Entrez	Source	Gene Description
3665	IRF7	■	■	■	■		■	■	■		■	■	s	interferon regulatory factor 7
5610	EIF2AK2	■	■	■			■	■	■			■	s	eukaryotic translation initiation factor 2-alpha kinase 2
3433	IFIT2	■	■	■		■	■	■				■	s	interferon-induced protein with tetratricopeptide repeats 2
6890	TAP1	■	■	■		■	■	■		■		■	s	transporter 1, ATP-binding cassette, sub-family B (MDR/TAP)
64135	IFIH1	■	■	■		■	■	■				■	s	interferon induced with helicase C domain 1
8638	OASL	■	■	■			■	■	■			■	s	2'-5'-oligoadenylate synthetase-like
6773	STAT2	■	■	■			■	■	■			■	s	signal transducer and activator of transcription 2, 113kDa
3669	ISG20	■	■	■			■	■	■			■	s	interferon stimulated exonuclease gene 20kDa
5696	PSMB8	■	■	■			■	■	■			■	s	proteasome (prosome, macropain) subunit, beta type, 8 (large multifunctional peptidase 7)
11274	USP18	■	■	■			■	■	■			■	s	ubiquitin specific peptidase 18
4599	MX1	■	■	■			■	■	■			■	s	myxovirus (influenza virus) resistance 1, interferon-inducible protein p78 (mouse)
3430	IFI35	■	■	■			■	■	■			■	s	interferon-induced protein 35
3437	IFIT3	■	■	■			■	■	■			■	s	interferon-induced protein with tetratricopeptide repeats 3
9246	UBE2L6	■	■	■		■	■	■				■	s	ubiquitin-conjugating enzyme E2L 6
5698	PSMB9	■	■	■			■	■	■			■	s	proteasome (prosome, macropain) subunit, beta type, 9 (large multifunctional peptidase 2)
79132	DHX58	■	■	■								■	s	DEXH (Asp-Glu-X-His) box polypeptide 58

How p-value is calculated

▸ Investigate Gene Sets
▸ View Gene Families
▸ Help

Converted 250 submitted identifiers into 196 entrez genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
CP	10	1329	196	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel | GenomeSpace

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ❓	FDR q-value ❓
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	50		$5.65 \, e^{-40}$	$7.51 \, e^{-37}$
REACTOME_INTERFERON_SIGNALING [159]	Genes involved in Interferon Signaling	26		$1.29 \, e^{-33}$	$8.6 \, e^{-31}$
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	30		$2.04 \, e^{-33}$	$9.04 \, e^{-31}$
REACTOME_INTERFERON_ALPHA_BETA_SIGNALI ALING [64]	Genes involved in Interferon alpha/beta signaling	17		$2.91 \, e^{-26}$	$9.67 \, e^{-24}$
REACTOME_INTERFERON_GAMMA_SIGNALING [63]	Genes involved in Interferon	13			

Fisher test table

	query	non-query
pathway	17	196-17
non-pathway	64-17	45956-64-196+17

Expected overlap if independent

196*64/45956 = 0.27

Notice: MSigDB p-values are biased

Converted 250 submitted identifiers into 196 entrez genes. click here for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Universe (N)
CP	10	1329	196	45956

Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant FDR q-values (< 0.05) and black indicates less significant FDR q-values (>= 0.05).

Save to: Excel | GenomeSpace

our N=12000

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	50		$5.65\ e^{-40}$	$7.51\ e^{-37}$
REACTOME_INTERFERON_SIGNALING [159]	Genes involved in Interferon Signaling	26		$1.29\ e^{-33}$	$8.6\ e^{-31}$
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	30		$2.04\ e^{-33}$	$9.04\ e^{-31}$
REACTOME_INTERFERON_ALPHA_BETA_SIGNALI ALING [64]	Genes involved in Interferon alpha/beta signaling	17		$2.91\ e^{-26}$	$9.67\ e^{-24}$
REACTOME_INTERFERON_GAMMA_SIGNALING [63]	Genes involved in Interferon	13			

Fisher test table

	query	non-query
pathway	17	196-17
non-pathway	64-17	45956-64-196+17

Expected overlap if independent

196*64/45956 = 0.27 ⟶ = 1.05

17

Looking into reactome pathways

Google: reactome interferon alpha beta signalling

Click to open in pathway browser

Let's look how the whole pathway looks in our dataset

Click on IFNa/b signalling

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p
REACTOME_IMMUNE_SYSTEM [933]	Genes involved in Immune System	50		
REACTOME_INTERFERON_SIGNALING [159]	Genes involved in Interferon Signaling	26		
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_NE_SYSTEM [270]	Genes involved in Cytokine Signaling in Immune system	30		
REACTOME_INTERFERON_ALPHA_BETA_SIGNALI ALING [64]	Genes involved in Interferon alpha/beta signaling	17		
REACTOME_INTERFERON_GAMMA_SIGNALING [63]	Genes involved in Interferon gamma signaling	13		
PID_IL23_PATHWAY [37]	IL23-mediated signaling events	8		
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERA ERACTION [267]	Cytokine-cytokine receptor interaction	14		
NABA_MATRISOME_ASSOCIATED [753]	Ensemble of genes encoding ECM-associated proteins including ECM-affiliated proteins, ECM regulators and secreted factors	20		
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEM HEMOKINES [57]	Genes involved in Chemokine receptors bind chemokines	8		
REACTOME_ADAPTIVE_IMMUNE_SYSTEM [539]	Genes involved in Adaptive Immune System	17		

Gene Set: REACTOME_INTERFERON_ALPHA_BETA_SIGNALING

Standard name	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING
Systematic name	M973
Brief description	Genes involved in Interferon alpha/beta signaling
Full description or abstract	
Collection	C2: curated gene sets CP:REACTOME: Reactome gene sets
Source publication	
Exact source	
Related gene sets	
External links	http://www.reactome.org/cgi-bin/eventbrowser_st_id?ST_ID=REACT_25162
Organism	Homo sapiens
Contributed by	Reactome
Source platform	EntrezGeneIds
Dataset references	
Download gene set	format: grp	text	gmt	gmx	xml
Compute overlaps ?	(show collections to investigate for overlap with this gene set)
Compendia expression profiles ?	Human tissue compendium (Novartis) NCI-60 cell lines (National Cancer Institute)
Advanced query	Further investigate these 64 genes
Gene families ?	Categorize these 64 genes by gene family
Show members	(show 65 members mapped to 64 genes)
Version history	3.1: First introduced

click "text"

19

Exercise

☑ Get list of genes in IFNa/b reponse pathway from QuickGO:

- https://www.ebi.ac.uk/QuickGO

Put the genes from the pathway into Phantasus

Fit to window

Sort by t

Most of the genes go up on LPS

Tools/Plots/GSEA plot

Select "t"

Check "vertical"
(or don't)

Annotate by
"treatment"
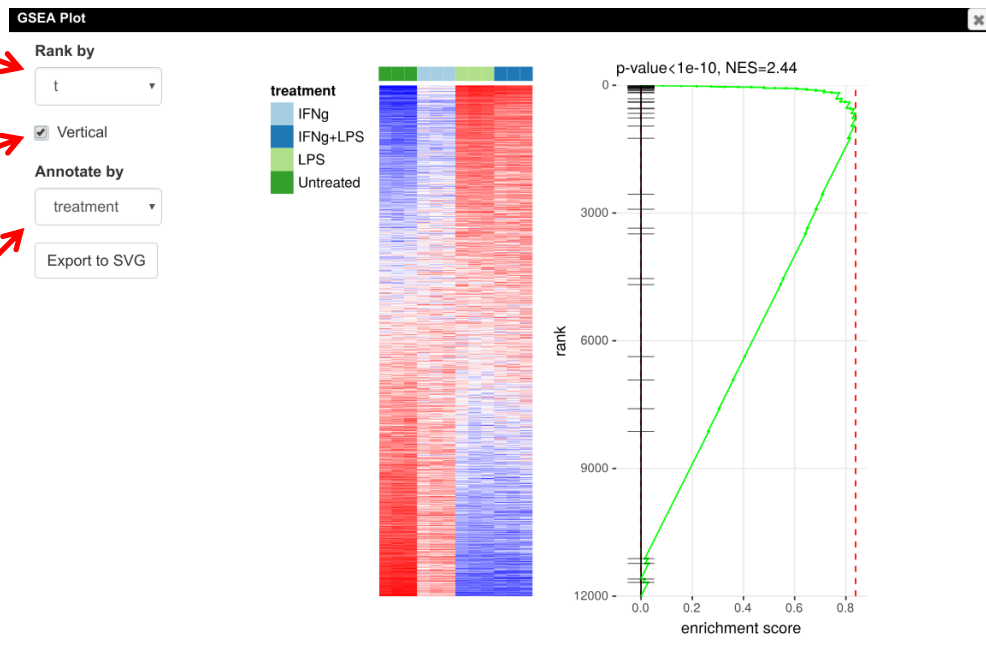
Pre-ranked gene set enrichment analysis

Enrichment
score function

ES(S)

Empirical null
distribution
from random sets

P-value & Normalized
Enrichment Score (NES)

http://www.ncbi.nlm.nih.gov/pubmed/16199517
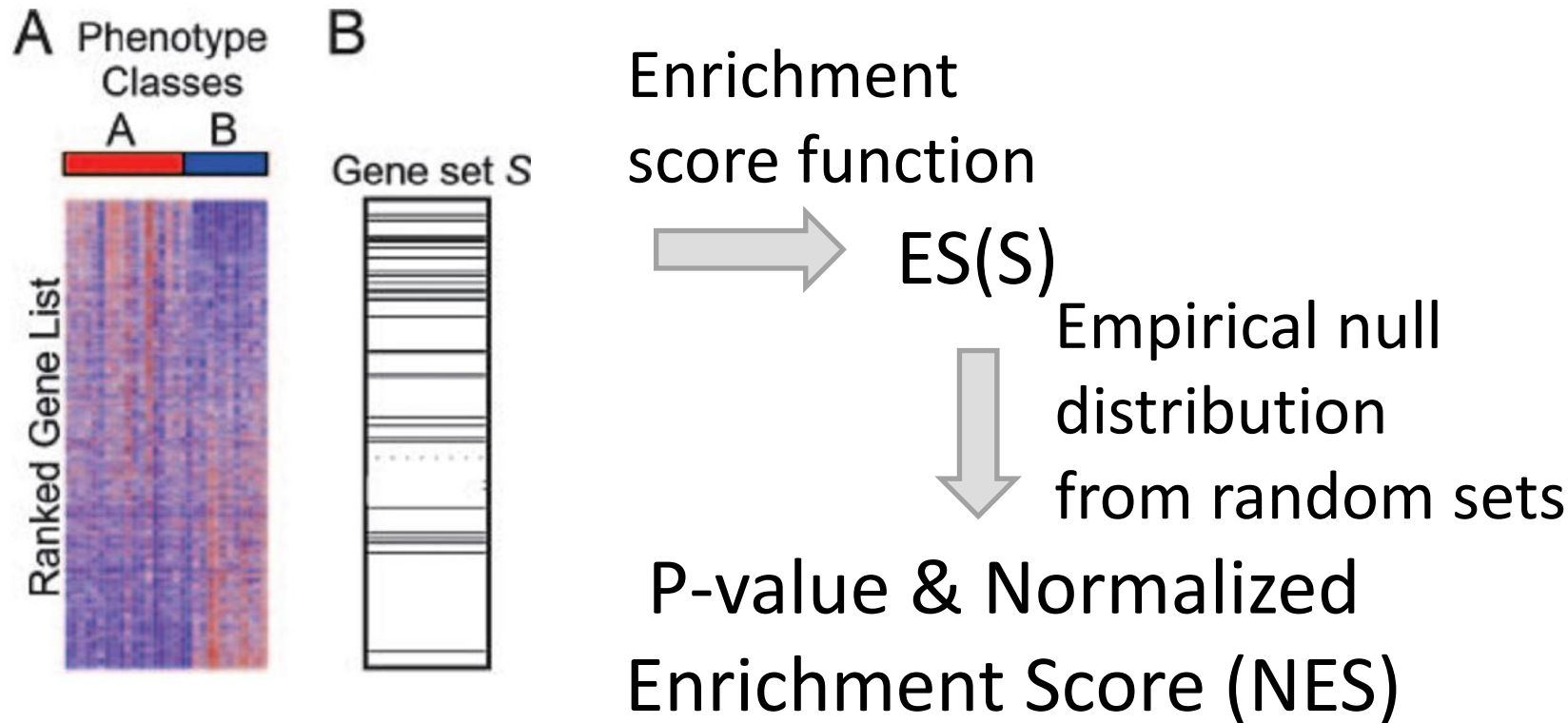
GSEA statistic calculation

Enrichment plot:
5991022_GLUCOSE_METABOLISM

Enrichment plot:
5991840_NEPHRIN_INTERACTIONS
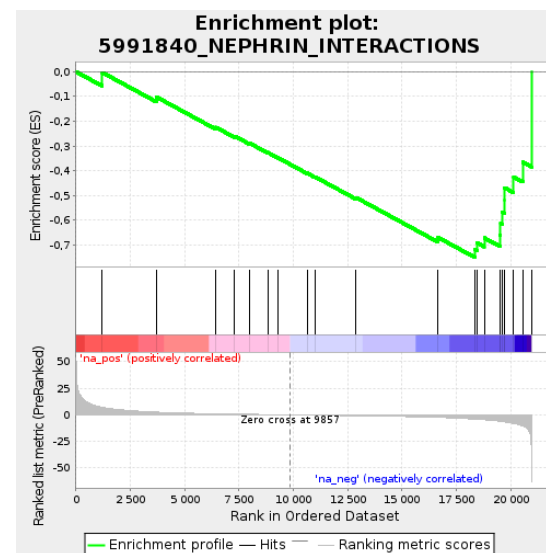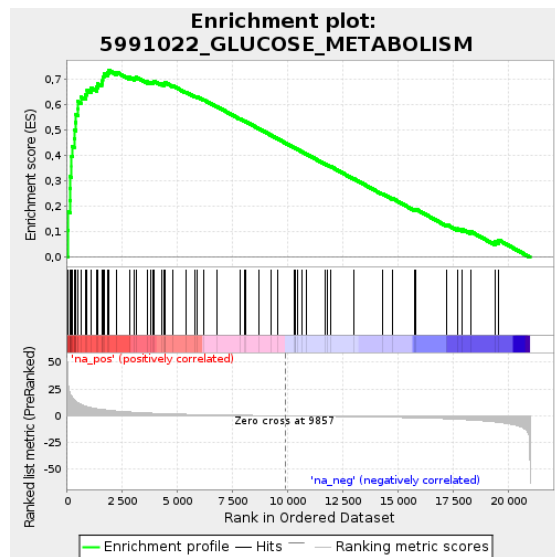
✔ ES[i] = ES[i-1] + |stat[i]| / NS, if $i \in$ gene set

✔ ES[i] = ES[i-1] − 1 / (N − k), if not

✔ NS = sum |stat[i]| for $i \in$ gene set

Fast gene set enrichment analysis (FGSEA)

✅ Fast estimation of arbitrarily low p-values

✅ https://github.com/ctlab/fgsea

✅ The error can be estimated

size	ES	Z	log2P (exact)	log2P (approx)	mean error	sd error	expected error
100	0.9	1001	-159.53812	-159.53769	0.00043054	0.61	0.48
100	0.7	1001	-79.54245	-79.5298	0.01265205	0.41	0.34
100	0.6	1001	-31.54746	-31.54093	0.00652617	0.26	0.21
100	0.5	1001	-13.56805	-13.5644	0.00364966	0.17	0.14

Gennady
Korotkevich

Vladimir
Sukhov

Running FGSEA

- ✔ Tools/Pathway analysis/Perform FGSEA
- ✔ Pathway database: MSigDB Hallmarks (Mouse)
- ✔ Rank by: t
- ✔ Column with gene ID: Gene ID

Actions:

Save as TSV

FGSEA:

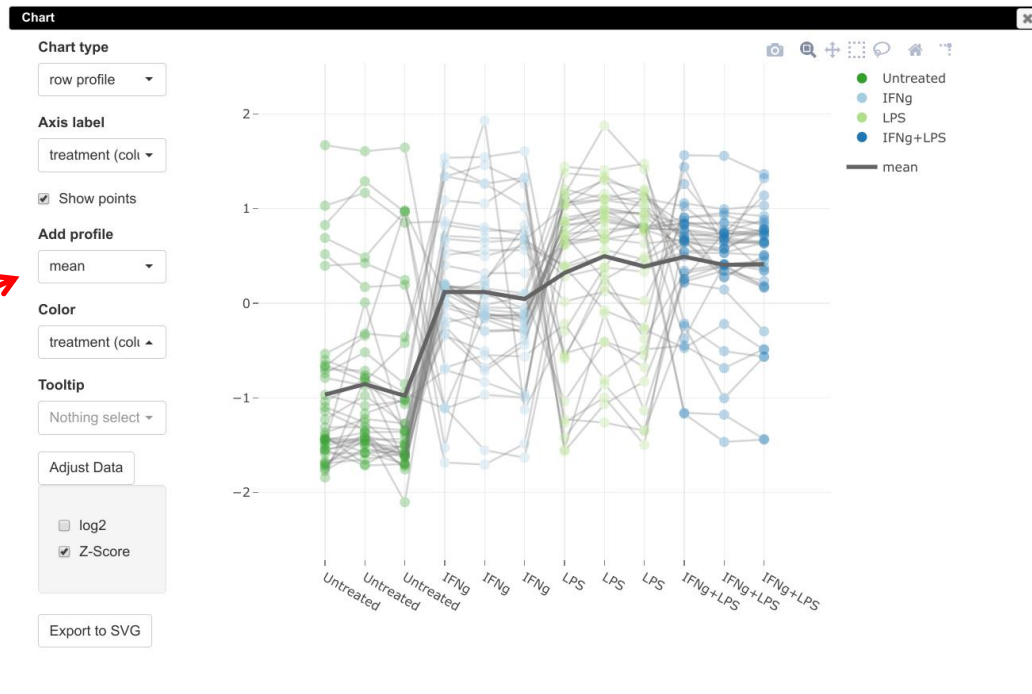| pathway	pval	padj	log2err	ES	NES	size	leadingEdge
Interferon gamma response	1.00e-10	8.17e-10	NA	0.836	3.14	183	17329 14962 20293 14293 16193 ...
Interferon alpha response	1.00e-10	8.17e-10	NA	0.814	2.82	94	15945 56066 99899 231655 57444 ...
Inflammatory response	1.00e-10	8.17e-10	NA	0.771	2.81	150	17329 17167 16175 20293 14293 ...
IL6-JAK-STAT3 signaling	1.00e-10	8.17e-10	NA	0.787	2.59	72	17329 330122 16193 12494 15945 ...
Allograft rejection	1.00e-10	8.17e-10	NA	0.681	2.47	142	17329 20293 16193 16176 18126 ...

Pathway detail:

Pathway name: Interferon gamma response
Pathway genes (ID): 56417 71939 71371 11861 3199 74 12010 242248 74481 12183 69550 12226 667277 317677 12362 12367 12363 12369 12370 20293 20304 20306 60533 12494 21939 12515 16149 12524 12575 14962 12628 12265 14747 22169 74157 12984 15945 56066 17329 230073 234311 80861 19106 66892 108670 14102 14129 14190 14293 55932 236573 14528 110168 14938 229003 67138 15251 14964 15006 15007 15013 15015 15018 110557 15040 630294 667977 14990 14991 14998 14960 14969 15894 15930 5

Pathway profile plot

✅ Tools/Chart

✅ Select genes from
   IFNa/b pathway

Add profile: mean
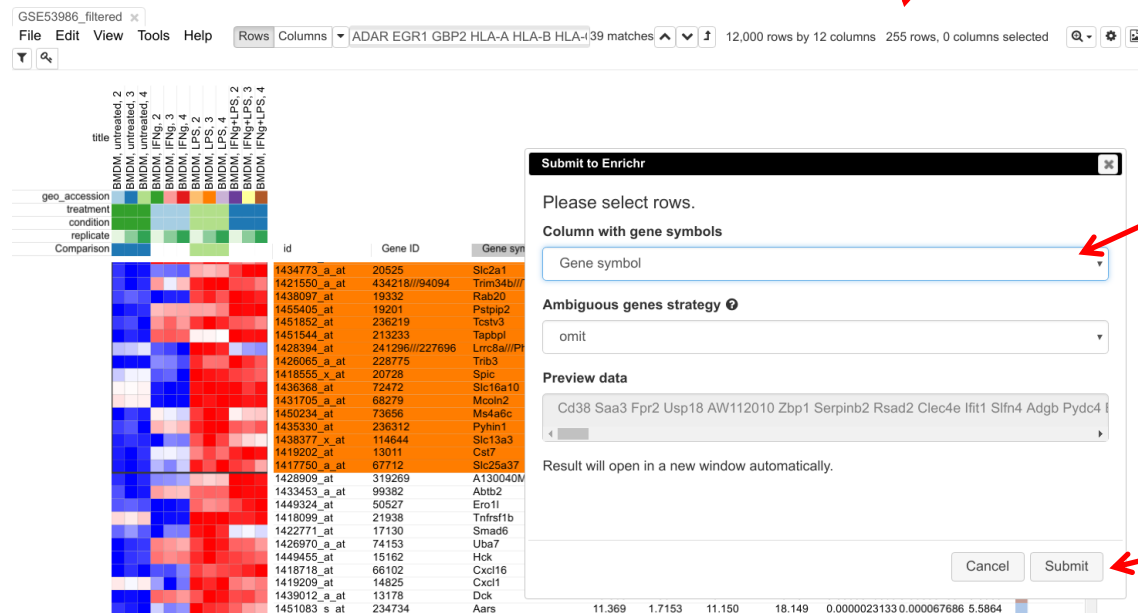
Adjust: Z-score

Using Enrichr

Select ~250 genes

✔ Tools/Pathway analsysis/Submit to Enrichr

Choose gene symbol column

Submit

28

Enrichr results

Check out different tabs
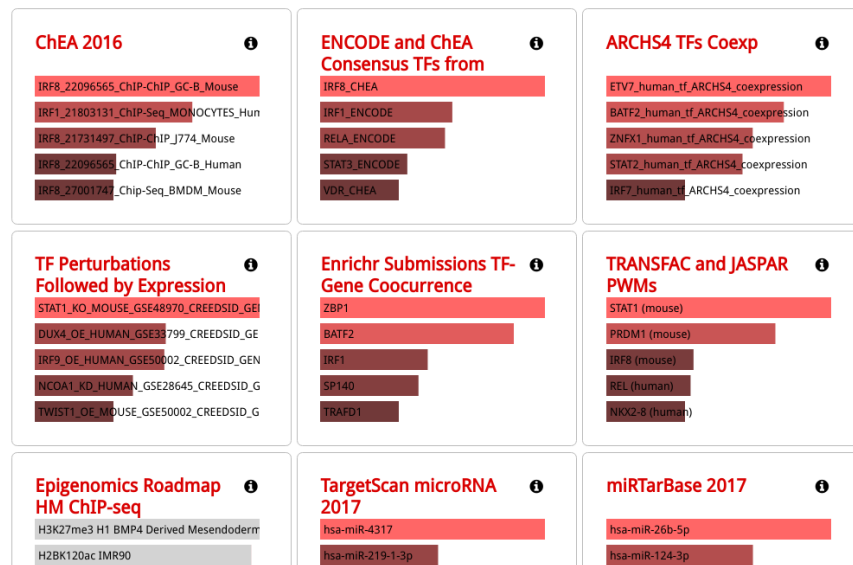
NB: if the new tab is not opening, probably pop-up windows are blocked

Other tools

- ✔ Gorilla
  - http://cbl-gorilla.cs.technion.ac.il/
- ✔ DAVID
  - https://david.ncifcrf.gov/home.jsp
- ✔ GSEA desktop tool
  - http://software.broadinstitute.org/gsea/downloads.jsp
- ✔ GenePattern
  - https://genepattern.broadinstitute.org/
- ✔ WebGestalt
  - http://www.webgestalt.org/
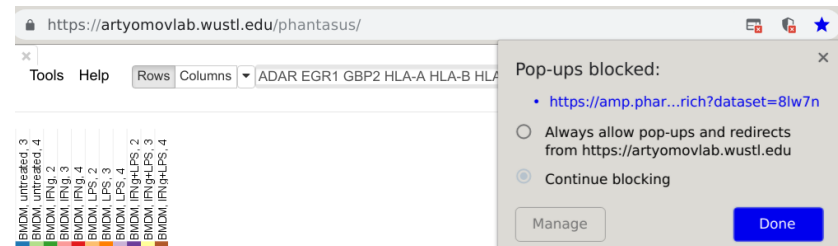
Summary

- ✅ Differential expression analysis and pathway enrichment are one of the most useful tools for data interpretation
- ✅ There are many online tools and databases for downstream analysis

- ✅ Next: working with public gene expression databases