Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

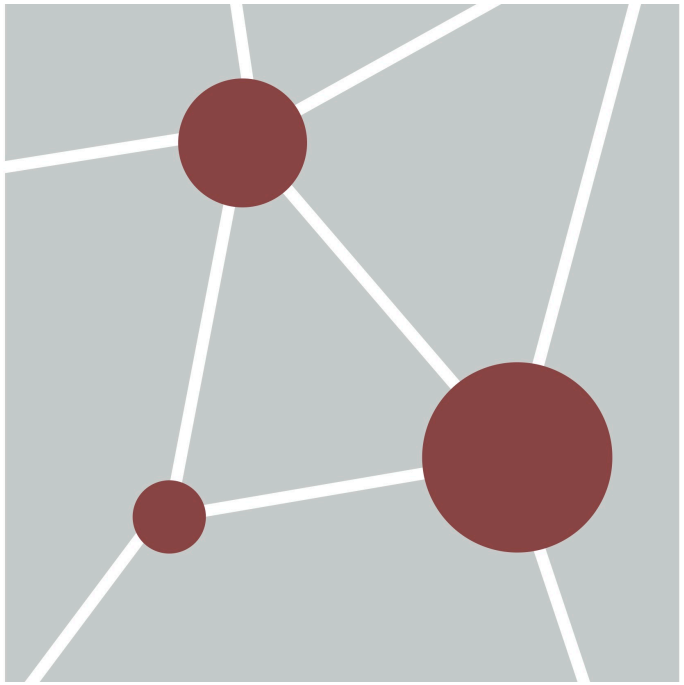
Finding over-represented pathways in gene lists

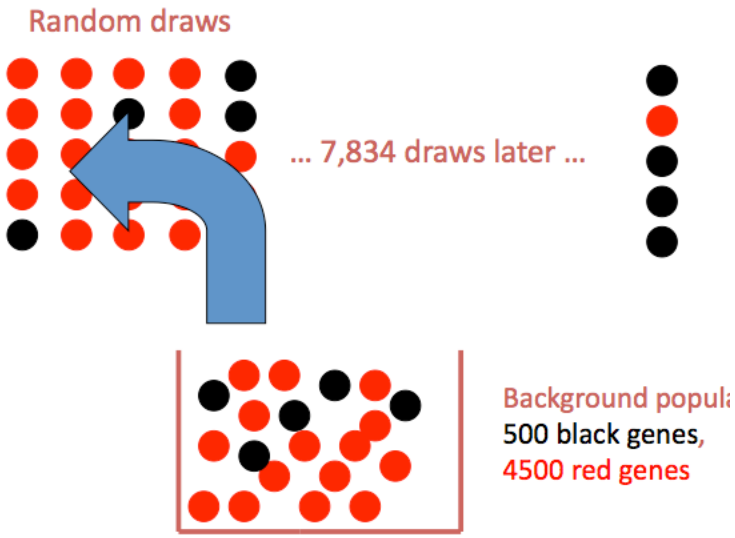
Veronique Voisin

Pathway and Network Analysis of –omics Data

June 26-28,2024

bioinformatics.ca



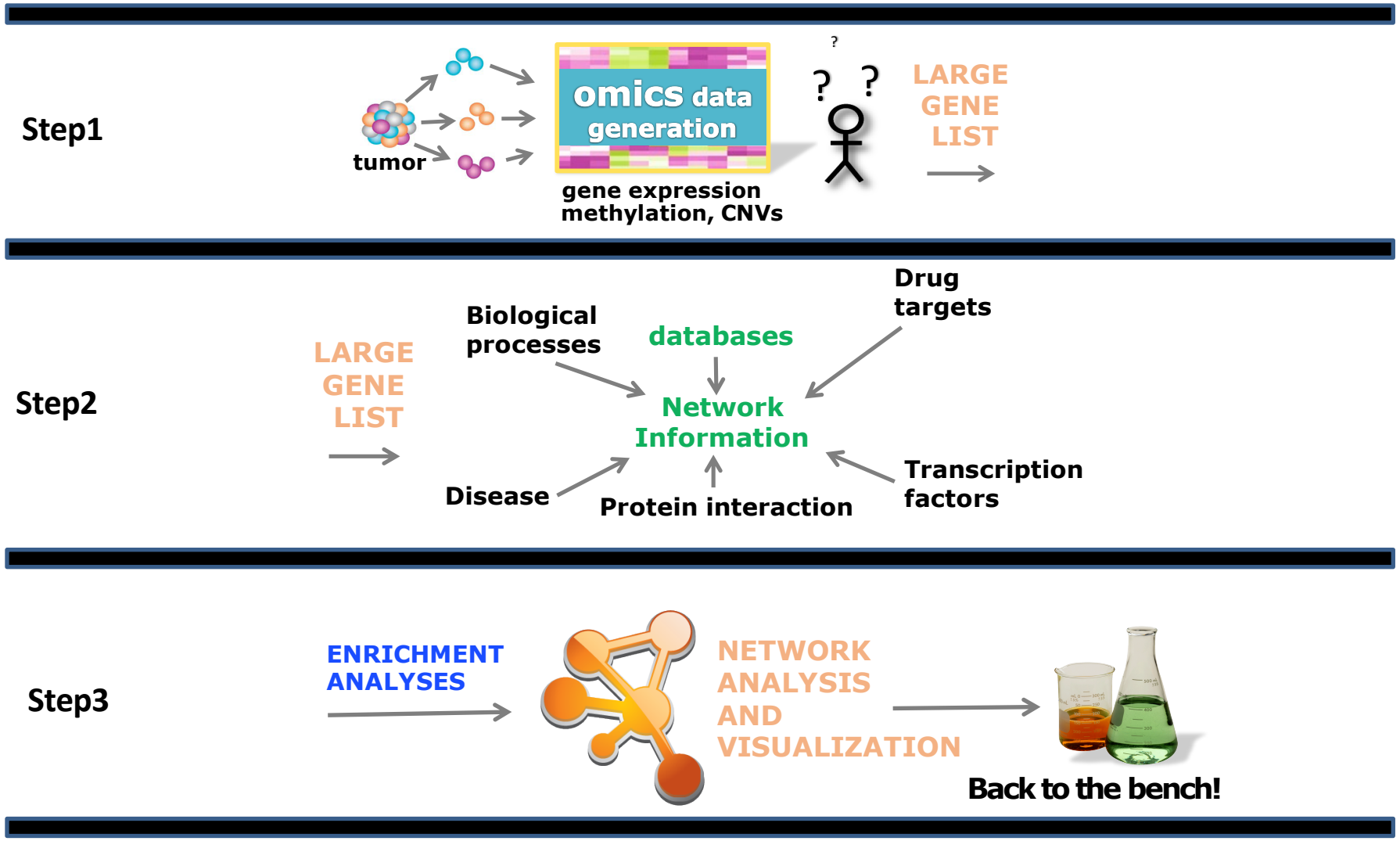


General Workflow of Enrichment Analysis and Definitions

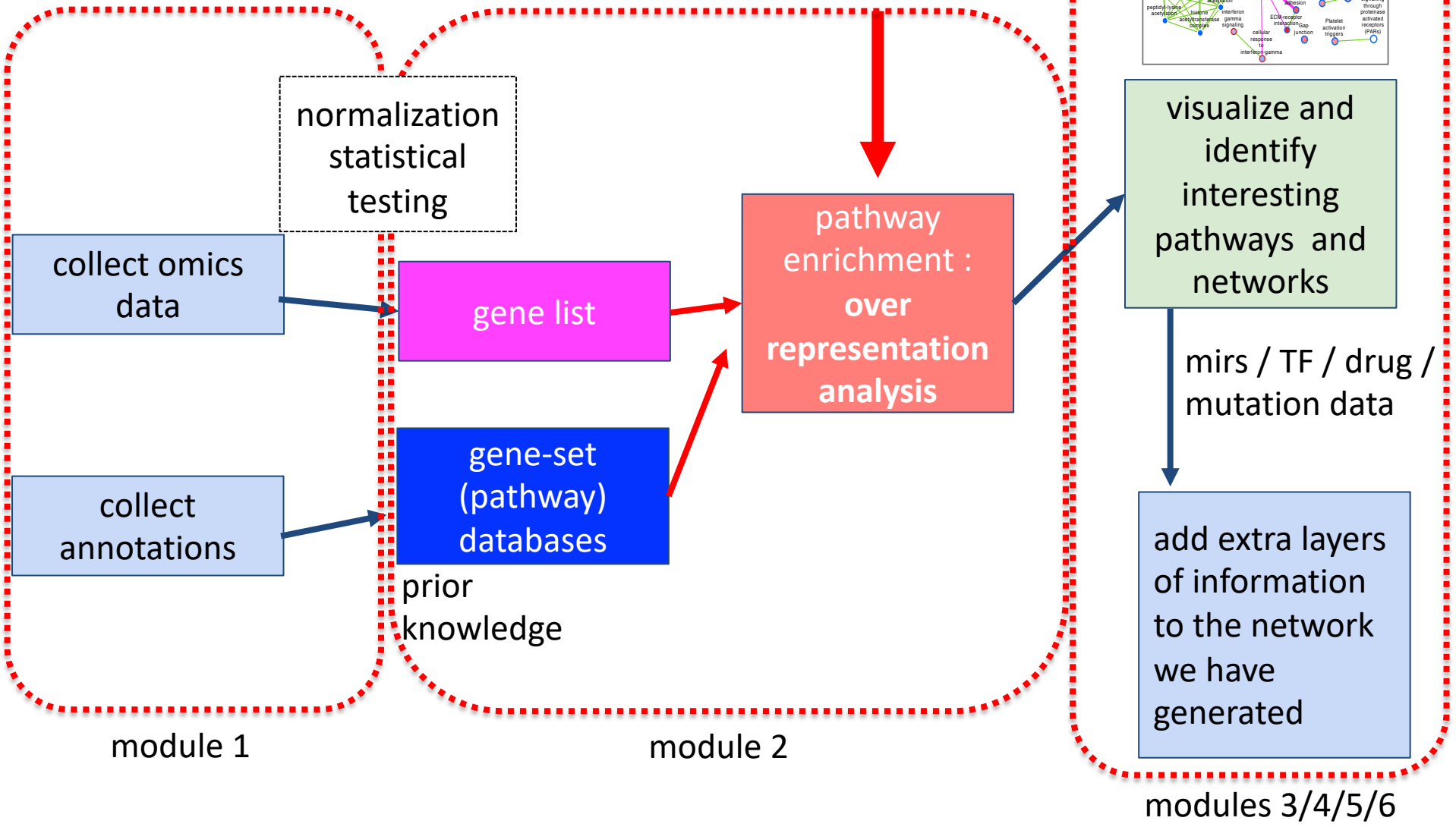
Learning Objectives

- Be able to understand the concept of overrepresentation analysis (ORA).
- Be able to understand the differences between a **defined gene list** and a **ranked gene list** and which enrichment test to apply.
- Be able to understand the concept of **pvalue** and **corrected pvalue (FDR)** in the context of enrichment analysis.
- Be able to understand the **result of an enrichment test** and how to interpret it
- Presentation of 2 enrichment tools

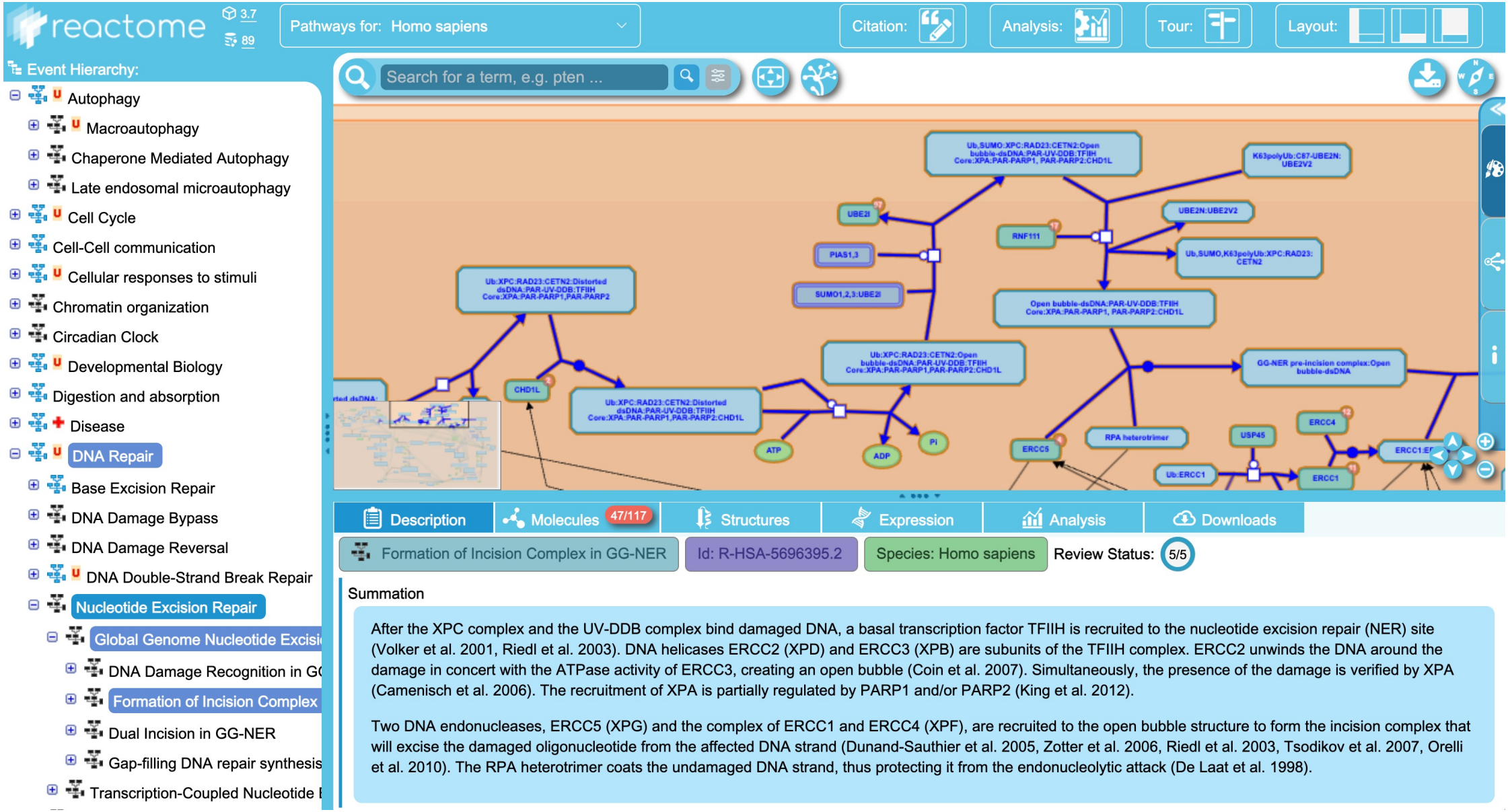
General Workflow of pathway and network analysis



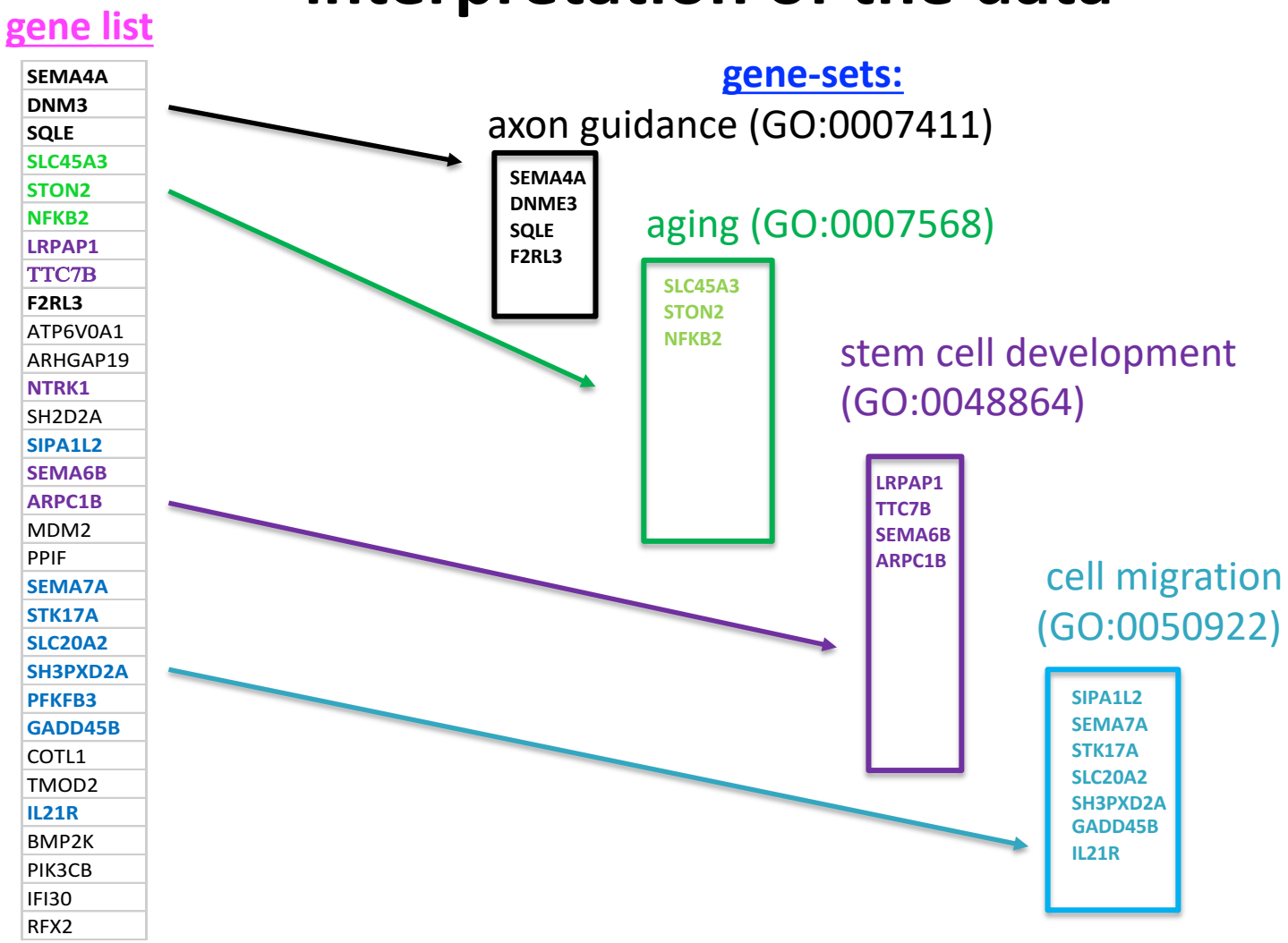
Analysis workflow



Pathway : annotated from database curator



Pathway enrichment analysis is a way to summarize your gene list into pathways to ease biological interpretation of the data



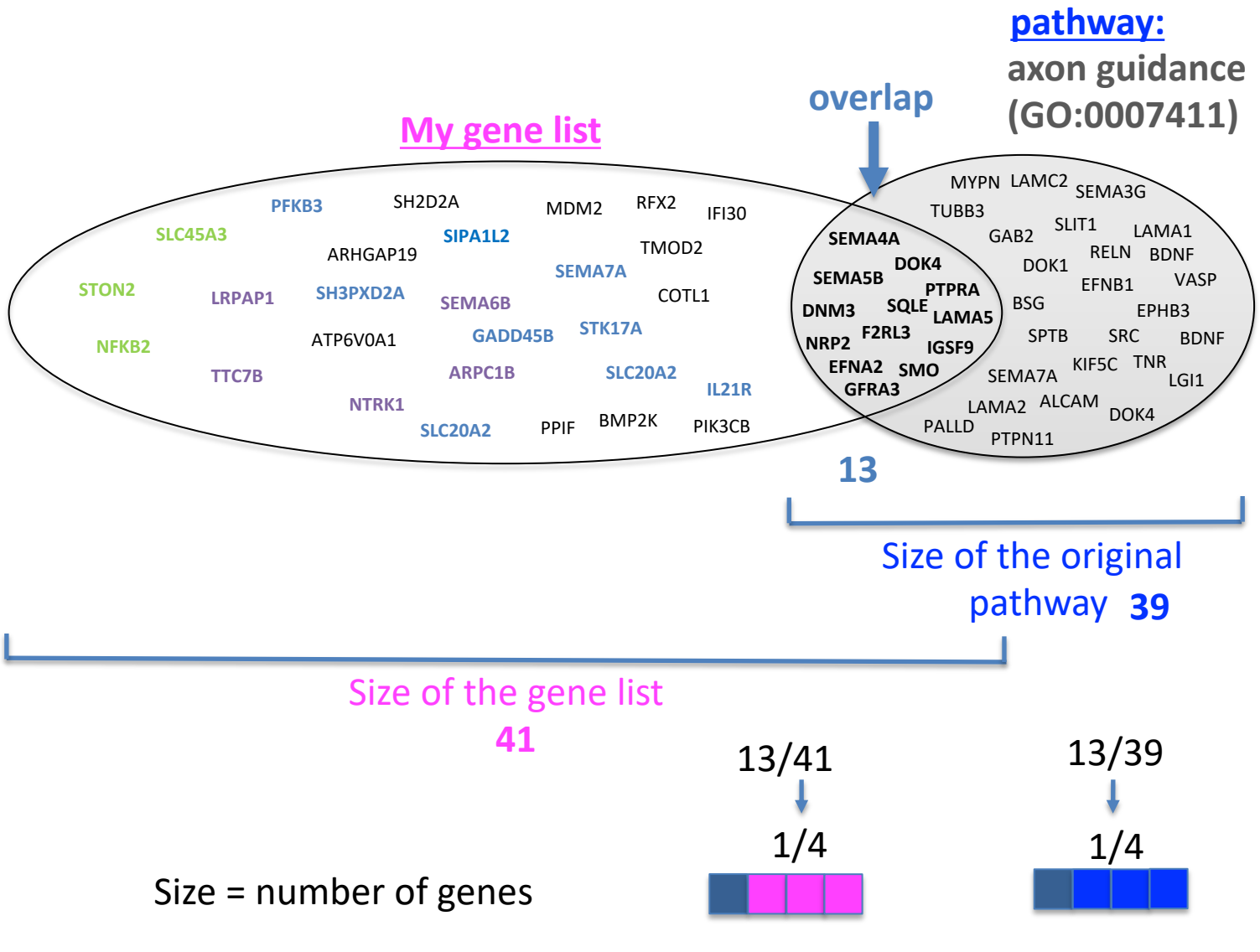
Pathway enrichment analysis calculates the overlap between our gene list and a pathway

gene list

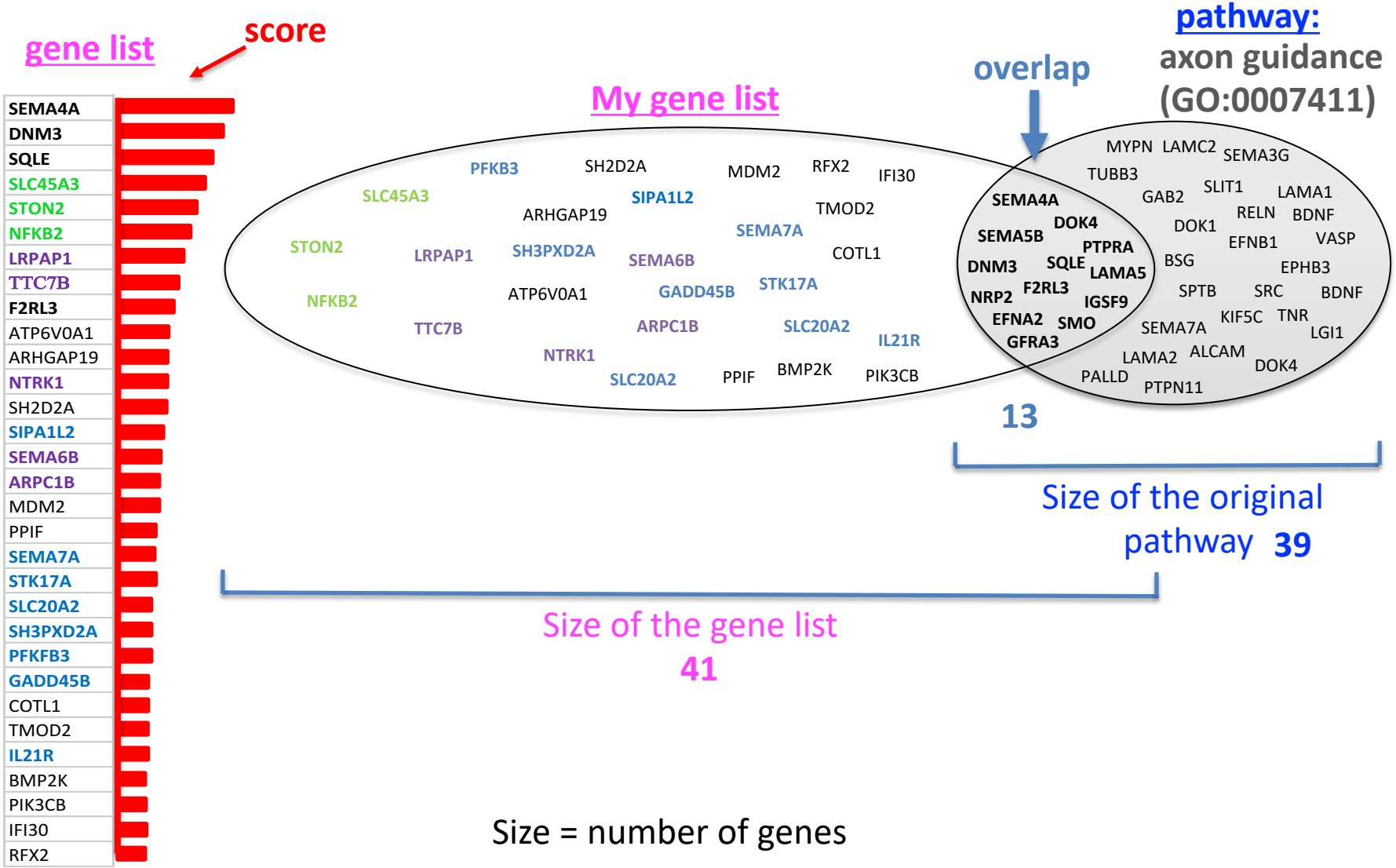
SEMA4A
DNM3
SQLE
SLC45A3
STON2
NFKB2
LRPAP1
TTC7B
F2RL3
ATP6V0A1
ARHGAP19
NTRK1
SH2D2A
SIPA1L2
SEMA6B
ARPC1B
MDM2
PPIF
SEMA7A
STK17A
SLC20A2
SH3PXD2A
PFKFB3
GADD45B
COTL1
TMOD2
IL21R
BMP2K
PIK3CB
IFI30
RFX2

• • •

FDR<0.05



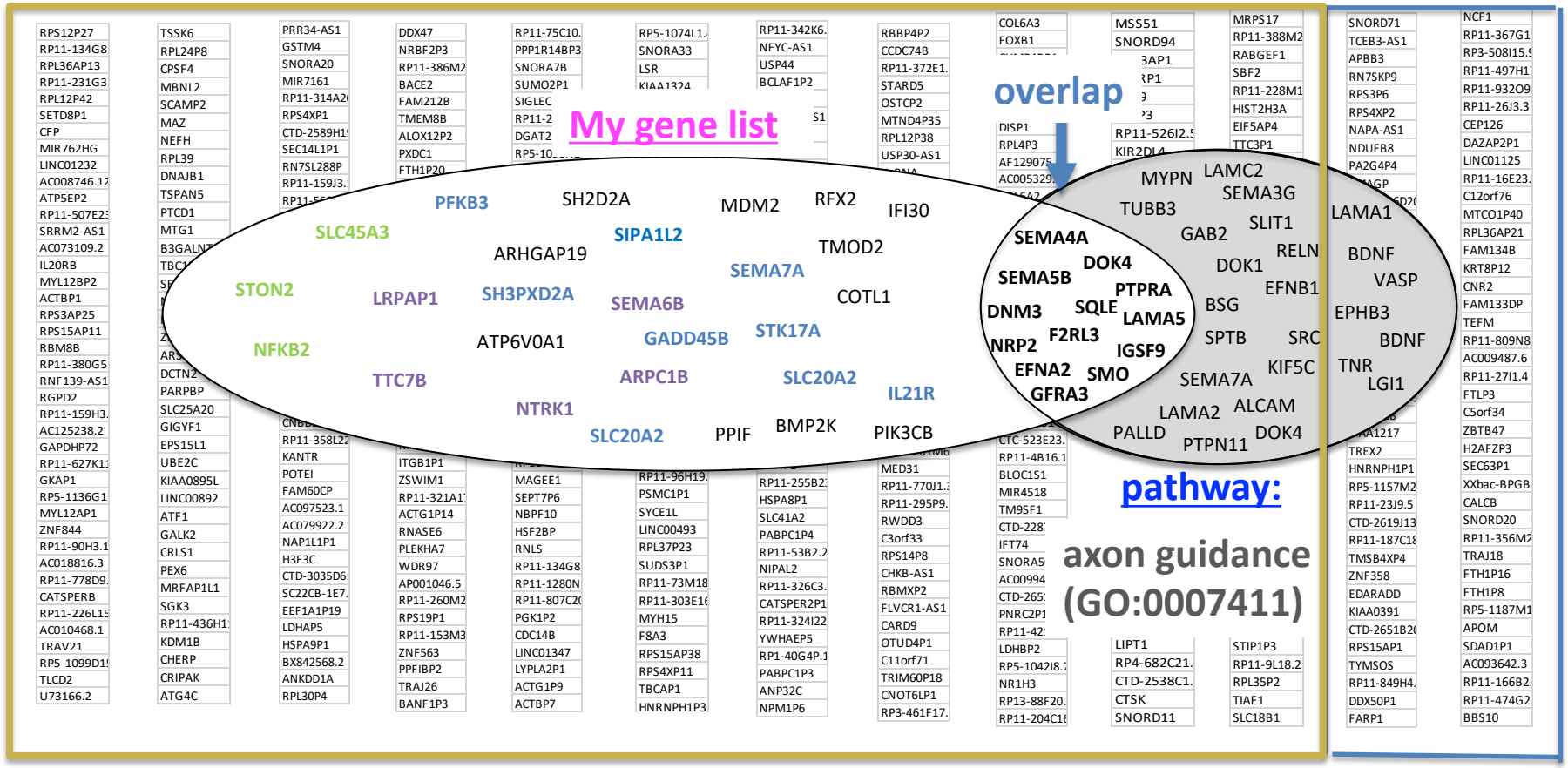
Can we add a score associated with the genes when calculating the enrichment score?



The background represents the genes that could have been captured in my omics experiment

genes measured in the experiment

genes not measured



estimated 20,000-25,000 **human** protein-**coding genes**

How many genes could have been captured in your experiment?

Over representation analysis

- The pathway is over-represented in our gene list.
- The pathway is enriched in our gene list.

Meaning:

- There are several or many genes from this pathway in our gene list.

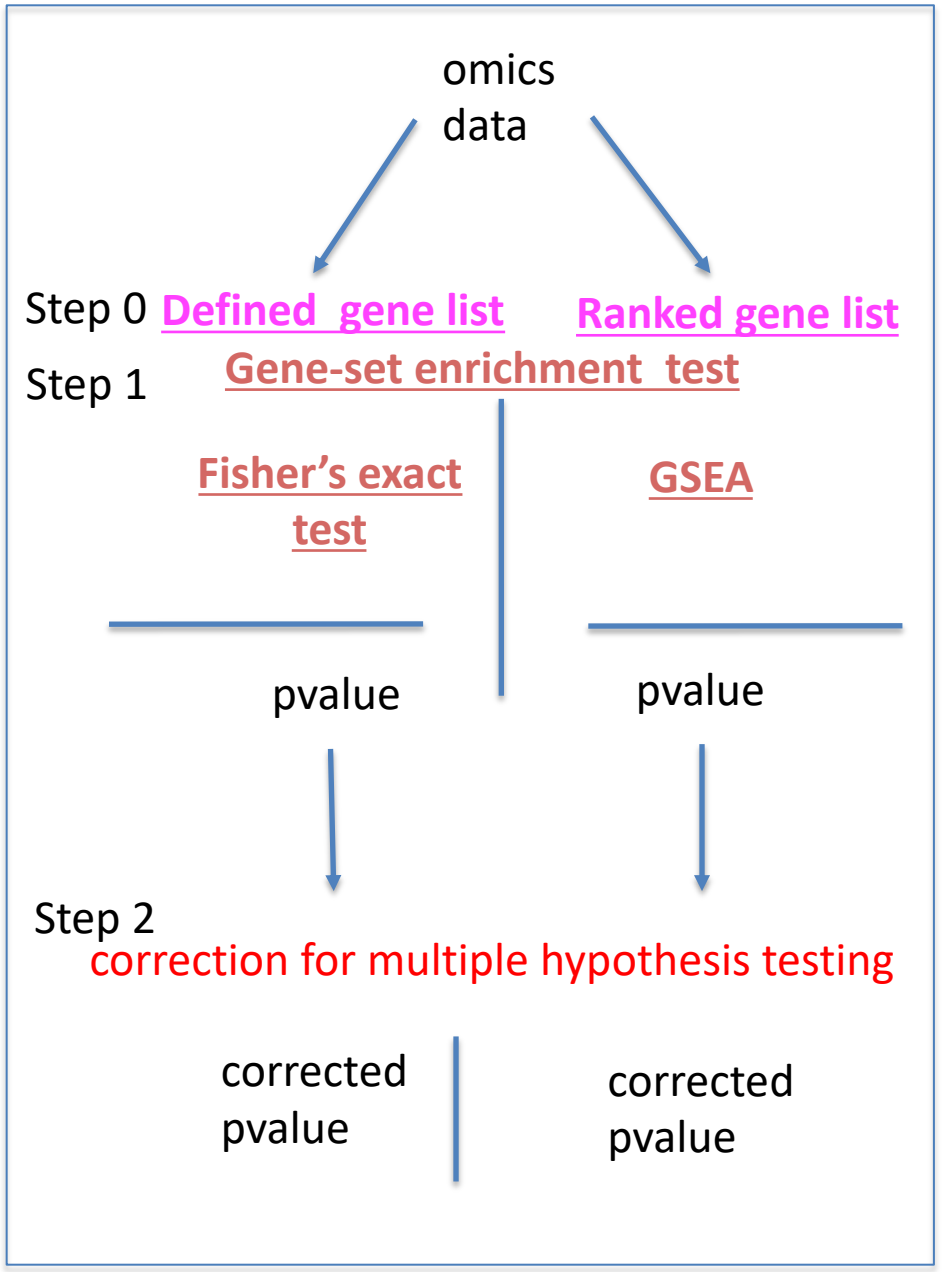
Definition:

- There are more genes from this pathway in our gene list than expected.
- There are more genes from this pathway in our gene list than what we could have obtained by chance only.

Enrichment Analysis using a Defined Gene List

Outline

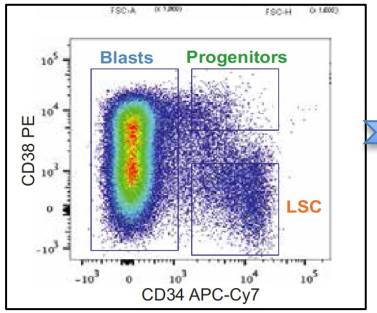
- Two types of gene lists (ranked or not)
- Introduction to enrichment analysis
- Fisher's Exact Test, aka Hypergeometric Test
- GSEA for ranked lists.
- Multiple test corrections:
  - Bonferroni correction
  - False Discovery Rate computation using Benjamini-Hochberg procedure

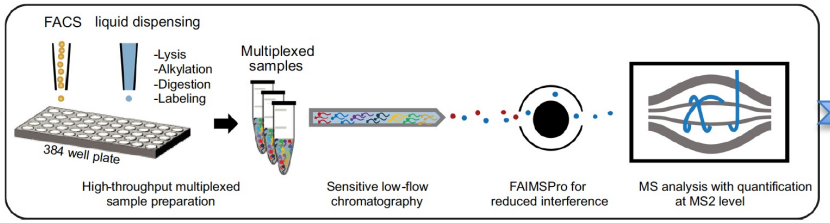


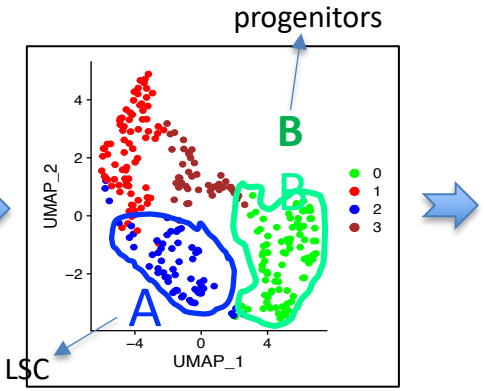
Types of enrichment analysis

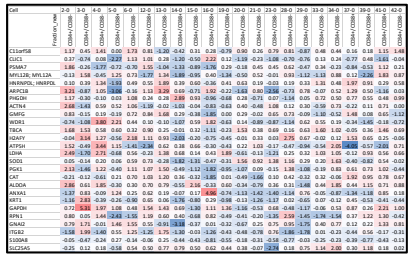
- **Defined gene list** (e.g. expression change > 2-fold)
  - Answers the question: **Are any pathways (gene sets) surprisingly enriched (or depleted) in my gene list?**
  - Statistical test: Fisher's Exact Test (aka Hypergeometric test)
- **Ranked gene list** (e.g. by differential expression)
  - Answers the question: **Are any pathways (gene sets) ranked surprisingly high or low in my ranked list of genes?**
  - Statistical test: **GSEA**, Wilcoxon rank sum test (+ others we won't discuss)

Example 1: enrichment analysis using a defined gene list

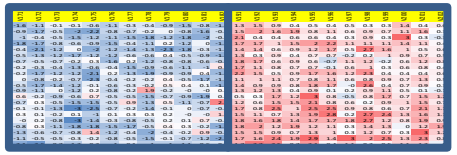








clusterA clusterB



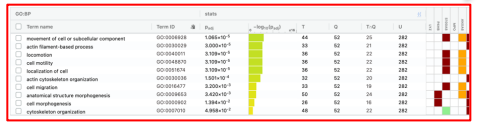
TOP 20 PROTEINS of CLUSTER B

LYZ
P4HB
S100A8
MPO
ANXA6
ACTN1
PGD
ANXA1
ANXA5
LCP1
TMSB4X
RAC2
GPI
CAP1
ANXA11
TKT
CORO1A
SERPINB1
NONO

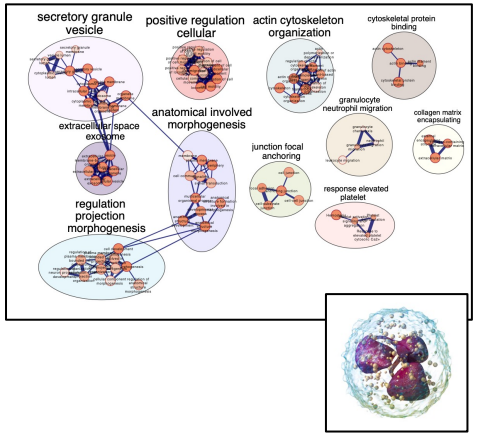
TOP 20 PROTEINS of CLUSTER A

TPM4
PRDX1
H2AFY
STMN1
RBMX
HMGB1
XRCC6
H2AFX
ANP32A
FSCN1
PTMA
C11orf58
HNRNPAB
HIST1H1C
TLN1
RPS27
HNRNPL
PRDX3
NUDC

g:Profiler



List of enriched pathways



- Sc-Proteomics: https://pubmed.ncbi.nlm.nih.gov/34099695/

What a pathway file looks like (showing only a few pathways):

Gene list

TOP 20 PROTEINS of CLUSTER B

LYZ
P4HB
S100A8
MPO
ANXA6
ACTN1
PGD
ANXA1
ANXA5
LCP1
TMSB4X
RAC2
GPI
CAP1
ANXA11
TKT
CORO1A
SERPINB1
NONO

Pathway file

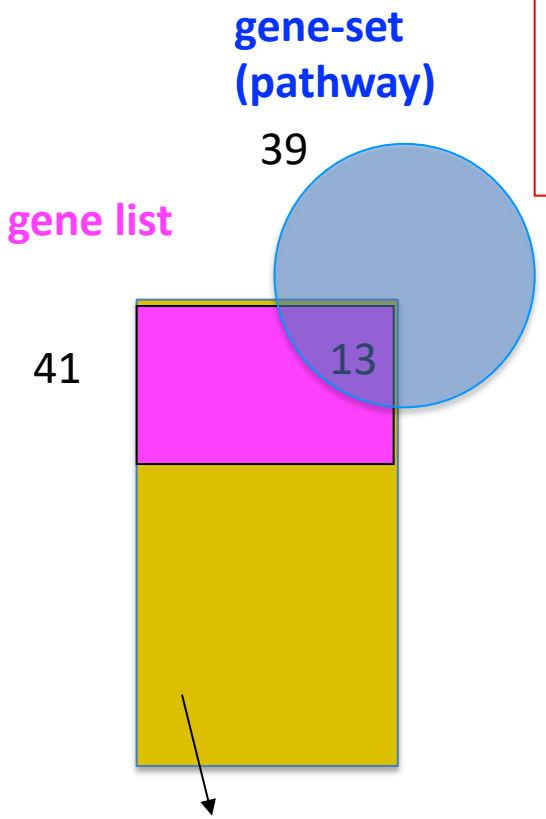
negative regulation of protein complex assembly	DACT1	ADD2	SOST	PFN1	ULK1	HEY2	MAPRE1	TUBB4A	TBCD
mesenchymal stem cell proliferation	SIX2								
interneuron axon guidance	LHX1	LHX9							
peptidyl-lysine oxidation	LOXL3	LOXL4	LOXL1	LOX	LOXL2				
negative regulation of skeletal muscle cell proliferation	EPHB1	MSTN	AKIRIN1						
tetrapyrrole catabolic process	HMOX1	BLVRB	UGT1A4	HMOX2	BLVRA	UGT1A1	AMBP		
regulation of adenylate cyclase-inhibiting adrenergic receptor signaling pathway	RGS2								
Purkinje myocyte action potential	SCN1B	TRPM4	SCN5A						
negative regulation of calcineurin-mediated signaling	FHL2	MYOZ1	HOMER2	MYOZ2	GSK3B	CHP1	RCAN1	PRNP	ACTN3
DNA endoreduplication	ZPR1								
protein maturation by protein folding	AIP	CALR	FKBP1A	CHCHD4	FKBP1B	WFS1			
regulation of histone H4-K16 acetylation	AUTS2	SMARCB1	PIH1D1	SIRT1	BRCA1				
sterol esterification	LCAT	SOAT1	ACAT1	SOAT2					
regulation of fat cell proliferation	TFDP1	GATA2	E2F1	E2F3	PID1	VSTM2A	PER2		
regulation of transcription from RNA polymerase II promoter in response to hypoxia	RBX1	NOTCH1	HIF1AN	HIGD1A	PSMD10	PSMD12	STOX1	RBPJ	CITED2
regulation of caveolin-mediated endocytosis	CLN3	PROM2	NEDD4L	SRC	UNC119				
blastocyst growth	ACVR1C	ZPR1							
desmosome assembly	PKP3	PRKCA	JUP	PKP2					
nuclear retention of unspliced pre-mRNA at the site of transcription	PRPF18	EXOSC10							
response to zinc ion	ATP13A2	MT1HL1	CRIP1	GLRA1	GLRA2	KHK	MT1DP	HVCN1	HAAO
immune response-activating signal transduction	HSP90AA1	KIR2DS2	TIRAP	TRAF3	TRAF6	CLEC4C	NOD2	EIF2B4	EIF2B3
negative regulation of very-low-density lipoprotein particle remodeling	APOA2	NR1H4	APOA1	APOC3					
rRNA processing	BMS1	HELB	RPUSD1	RPUSD2	RPL7L1	UTP11	POP5	WDR43	NOLC1
specification of mesonephric tubule identity	OSR1								
regulation of heart induction	GATA5	DKK1	ROBO2	ROBO1	MESP1	WNT3A			
lateral attachment of mitotic spindle microtubules to kinetochore	CENPE								
histone H2B ubiquitination	DTX3L	LEO1	RNF8	RNF40	UBE2E1	RNF20	WAC	PAF1	CTR9
monoacylglycerol biosynthetic process	MOGAT3	MOGAT2	PLA2G4A	DGAT2L6	AWAT2	DGAT2	DGAT1		
hard palate development	FOXE1	SOX11							
positive regulation of wound healing	F2R	INSL3	SERPINE1	FERMT2	APOH	SMOC2	OCLN	HRAS	PLEK
thalamus development	LRP6	PTCHD1	CHRNB2	CNTNAP2					
dendritic cell homeostasis	GPR183								
positive regulation of trophectodermal cell proliferation	IGF1								
positive regulation of nuclear-transcribed mRNA catabolic process, deadenylation-depen	CPEB3	NANOS1	NANOS2	CNOT7	CNOT1	POLR2G	DHX36	NANOS3	TNRC6C

18000 pathways in file

Yellow: proteins that match with my list

Overlap between my gene list and pathway = 4

How do simple enrichment tests work?



OVERLAP BETWEEN GENE LIST AND GENE_SETS (PATHWAYS)

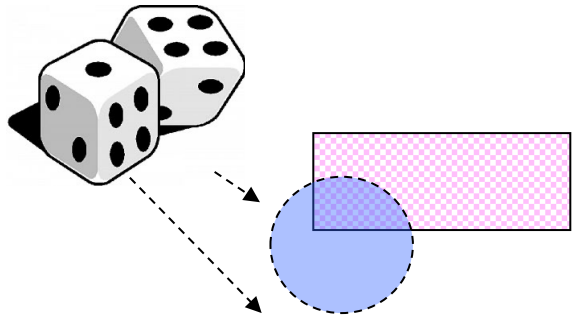
1

Enrichment Table

Gene-set	p-value
Spindle	0.0001
Apoptosis	0.025

2

Is this overlap larger than expected by chance?



random sampling using background genes

Empirical pval =(#obs_overlap > random_overlap) +1) / (number of tests + 1)

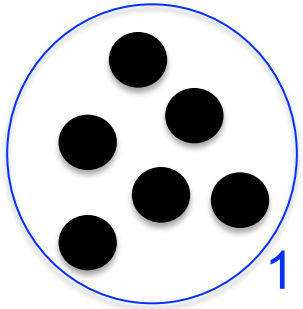
The Fisher's exact test

a.k.a., hypergeometric test

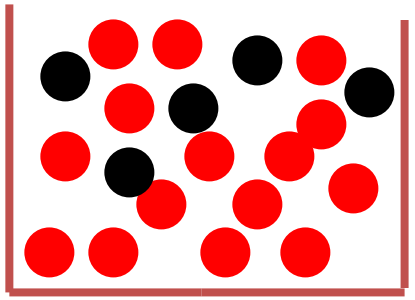
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

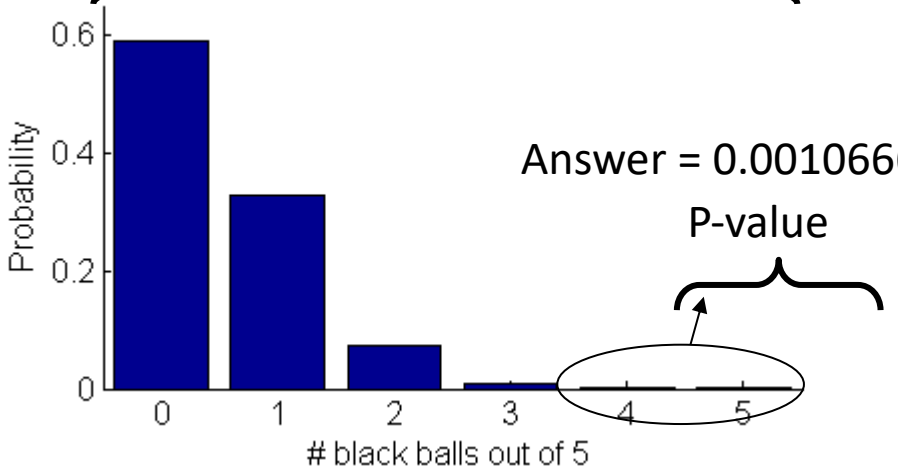
Background population:
500 black genes,
4500 red genes



1 gene-set/pathway

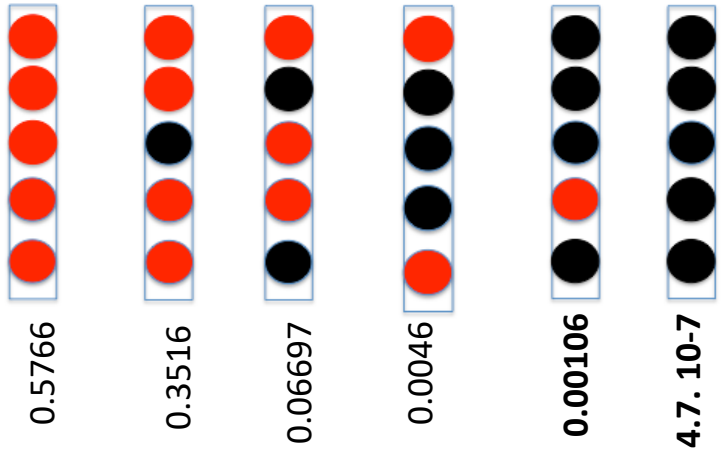


Null distribution



Answer = 0.001066662

P-value

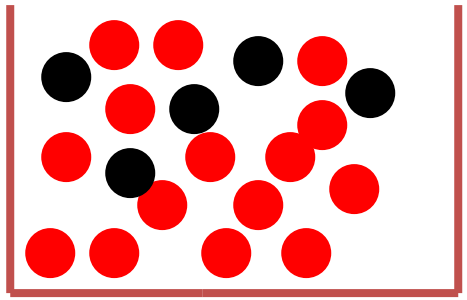


hypergeometric probability distribution

2x2 contingency table for Fisher's Exact Test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



	In gene list	Not in gene list	
In pathway	x = 4	496	m = 500
Not in pathway	k-x = 1	4499	t – m = 4500
	k= 5	4995	t =5000

$$P(X = x > q) = \sum_{x=q}^{m} \frac{\binom{m}{x}\binom{t-m}{k-x}}{\binom{t}{k}}.$$

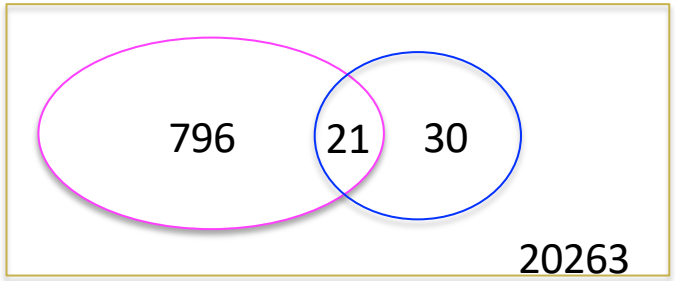
Background population:
500 black genes,
4500 red genes

g:Profiler

GO:BP | stats

Term name	Term ID	pvalue	$-\log_{10}(p_{adj})$	T	Q	T∩Q	U
actin filament-based process	GO:0030029	2.2e-16		817	51	21	21110
regulation of actin filament-based process	GO:0032970	4.9e-08		408	51	10	21110
regulation of cytoskeleton organization	GO:0051493	6.474e-07		543	51	10	21110
organelle localization	GO:0051640	1.769e-05		583	51	10	21110

T (term): pathway that is being tested
Q (query): my gene list
TnQ: overlap between pathway and gene list
U (universe): background



796+21=817
30+21-21
21110-30-21-796=20263

Possibility to upload a custom background:
Let's say we only could measure 4000 proteins: redo the calculation and see how it affects the results.

2x2 contingency table

	In protein list	Not in protein list
In pathway	21	796
Not in pathway	30	20263

Fisher's exact test

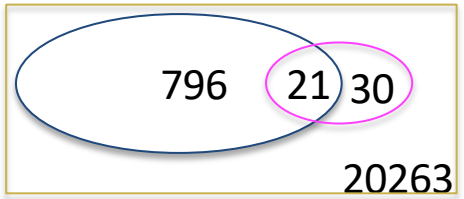
Total of GO:BP pathways tested is 349

GO:BP

Term name	Term ID	pvalue	$-\log_{10}(p_{adj})$	T	Q	T∩Q	U
actin filament-based process	GO:0030029	2.2e-16		817	51	21	21110
regulation of actin filament-based process	GO:0032970	4.9e-08		408	51	10	21110
regulation of cytoskeleton organization	GO:0051493	6.474e-07		543	51	10	21110
organelle localization	GO:0051640	1.769e-05		583	51	10	21110

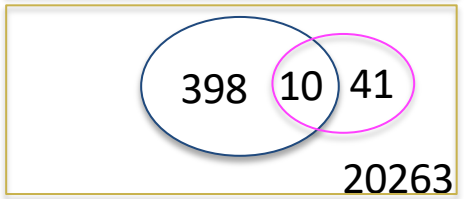
P value

Actin filament-based process



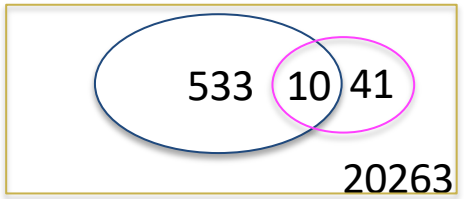
2.2e-16 (0.00000000000000022)

Regulation of actin filament-based process



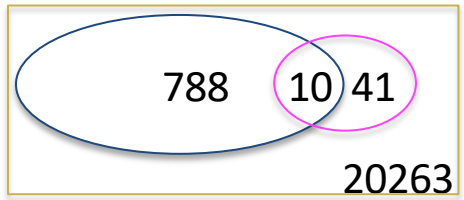
4.9e-08 (0.000000049)

Regulation of cytoskeleton organization



6.474e-07 (0.000000647)

Organelle localization



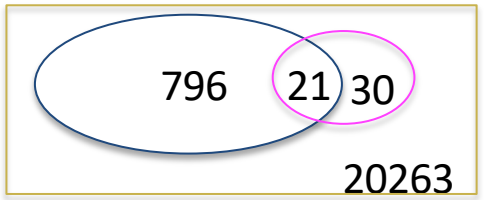
1.769e-05 (0.00001769)

Total of GO:BP pathways tested is 349

GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$	T	Q	TnQ	U
actin filament-based process	GO:0030029	2.2e-16		817	51	21	21110
regulation of actin filament-based process	GO:0032970	4.9e-08		408	51	10	21110
regulation of cytoskeleton organization	GO:0051493	6.474e-07		543	51	10	21110
organelle localization	GO:0051640	1.769e-05		583	51	10	21110
endocytosis	GO:0006897			798	51	10	21110
entry into host	GO:0044409			159	51	2	21110
monoatomic ion homeostasis	GO:0050801			602	51	4	21110
proton transmembrane transport	GO:1902600			161	51	2	21110
negative regulation of binding	GO:0051100	0.05875		161	51	2	21110
melanosome localization	GO:0032400	0.06087		25	51	1	21110

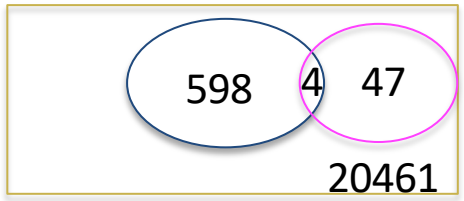
P value

Actin filament-based process



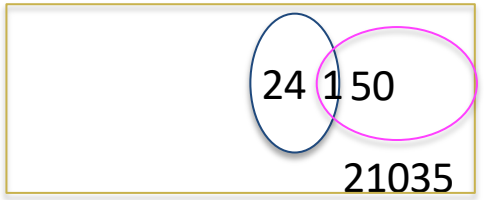
2.2e-16 (0.00000000000000022)

Monoatomic ion homeostasis



0.05875 Not significant

Melanosome localization

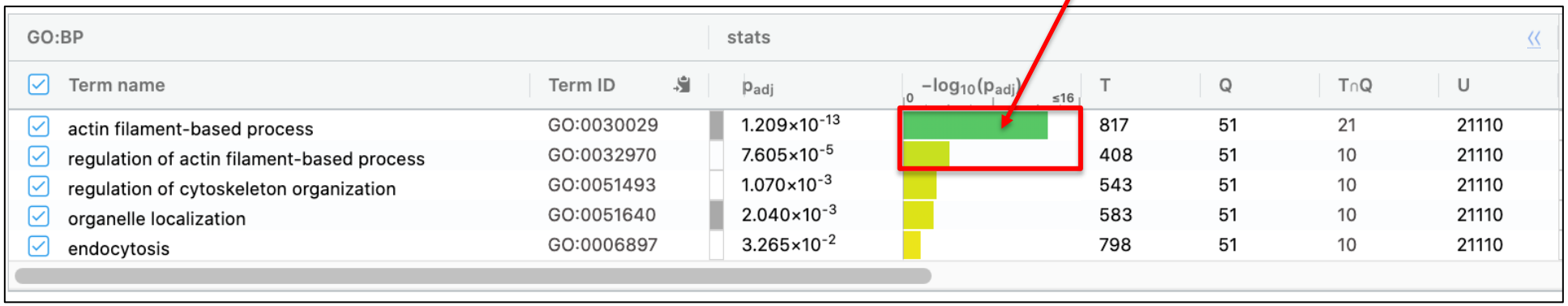


0.06087 Not significant

The p value assesses the probability that the tested pathway is enriched in our gene list by chance only.

We are testing many pathways at the same time

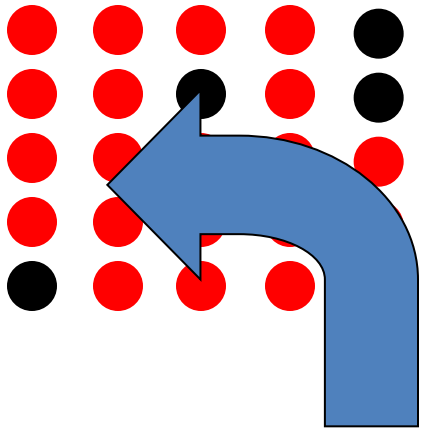
correction for multiple hypothesis testing



GO:BP		stats					
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ (0 … ≤16)	T	Q	T∩Q	U
actin filament-based process	GO:0030029	1.209×10^{-13}		817	51	21	21110
regulation of actin filament-based process	GO:0032970	7.605×10^{-5}		408	51	10	21110
regulation of cytoskeleton organization	GO:0051493	1.070×10^{-3}		543	51	10	21110
organelle localization	GO:0051640	2.040×10^{-3}		583	51	10	21110
endocytosis	GO:0006897	3.265×10^{-2}		798	51	10	21110

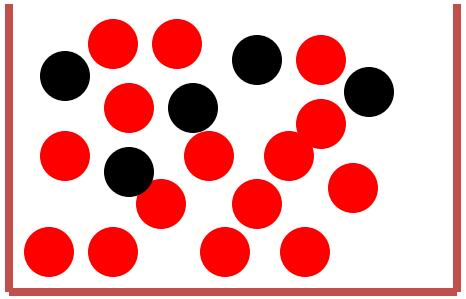
How to win the p-value lottery

Random draws

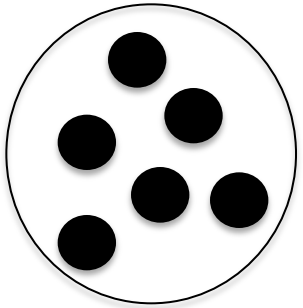


… 7,834 draws later …

Expect a random draw with observed enrichment once every 1 / P-value draws



Background population:
500 black genes,
4500 red genes



1 gene-set
(endocytosis)

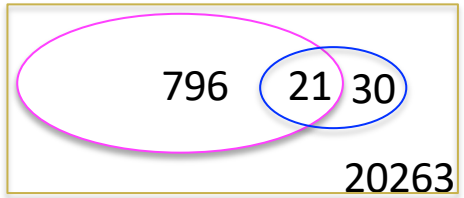
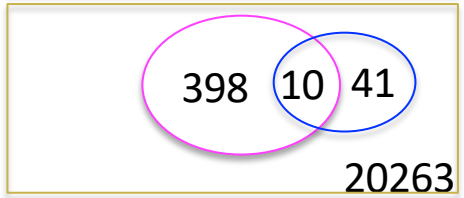
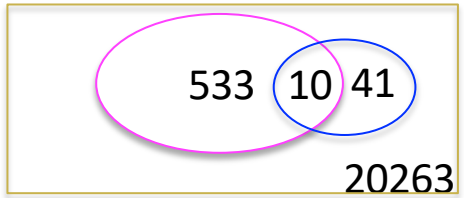
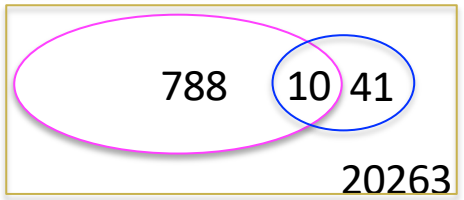
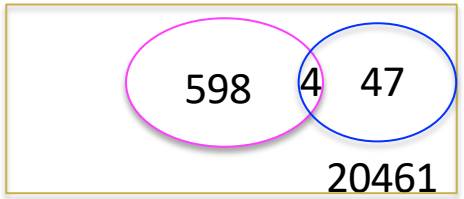
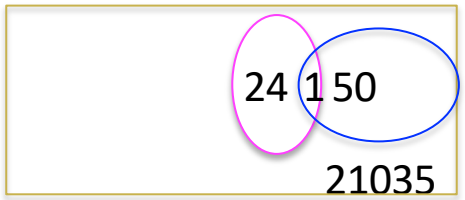
Simple P-value correction: Bonferroni

If *M* = # of gene-sets (pathways) tested:

Corrected P-value = *M* x original P-value

Corrected P-value is greater than or equal to the probability that **one or more** of the observed enrichments could be due to random draws. The jargon for this correction is **"controlling for the *Family-Wise Error Rate (FWER)*"**

Total of GO:BP pathways tested is 349

		P value	Adj. P value
Actin filament-based process		2.2e-16 * 349 =	7.678e-14
Regulation of actin filament-based process		4.9e-08 * 349 =	1.7101e-05
Regulation of cytoskeleton organization		6.474e-07 * 349 =	0.0002
Organelle localization		1.769e-05 * 349 =	0.006
Monoatomic ion homeostasis		0.05875 * 349	> 1
Melanosome localization		0.06087 * 349	> 1

False discovery rate (FDR)

- FDR is *the expected **proportion** of the observed enrichments due to random chance.*
- Compare to Bonferroni correction which is a bound on *the probability that **any one** of the observed enrichments could be due to random chance.*
- Typically FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the “q-value”

Benjamini-Hochberg example

Rank	Category	(Nominal) P-value	Adjusted P-value			FDR / Q-value
1.	Actin filament-based process	2.2e-16	2.2e-16	* 349/1	= 7.678e-14.	7.678e-14
2	Regulation of actin filament-based process	4.9e-08	4.9e-08	* 349/2	= 1.7101e-05	1.7101e-05
3	Regulation of cytoskeleton organization	6.474e-07	6.474e-07	* 349/3	= 0.0002	0.0002
4	Organelle localization	1.769e-05	1.769e-05	* 349/4	= 0.006	0.006
	...	....	....	....	....	....
347	ion homeostas	0.05870	0.05870	* 349/347	= 0.0590	0.0589
348	Monoatomic ion homeostasis	0.05875	0.05875	* 349/348	= 0.0589	0.0589
349	Melanosome localization	0.06087	0.06087	* 349/349	= 0.06	0.06

1. Sort P-values
of all tests in increasing order

2. Calculate Adjusted P-value :
P-value x [# of tests] / Rank

3. Calculate the Q-value (or FDR).

Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.

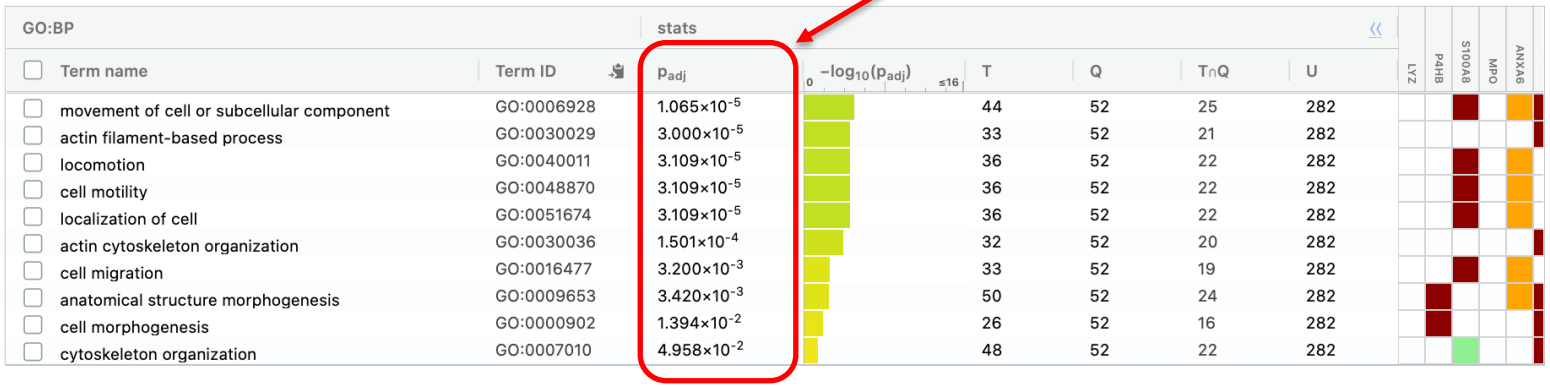
Select pathways significant at FDR < 0.05 for your analysis

g:Profiler

Significance threshold

Benjamini-Hochberg FDR

Padj = q value



GO:BP		stats									
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16	T	Q	TnQ	U	LYZ	P4HB	S100A8	MPO	ANXA6
movement of cell or subcellular component	GO:0006928	1.065×10^{-5}		44	52	25	282					
actin filament-based process	GO:0030029	3.000×10^{-5}		33	52	21	282					
locomotion	GO:0040011	3.109×10^{-5}		36	52	22	282					
cell motility	GO:0048870	3.109×10^{-5}		36	52	22	282					
localization of cell	GO:0051674	3.109×10^{-5}		36	52	22	282					
actin cytoskeleton organization	GO:0030036	1.501×10^{-4}		32	52	20	282					
cell migration	GO:0016477	3.200×10^{-3}		33	52	19	282					
anatomical structure morphogenesis	GO:0009653	3.420×10^{-3}		50	52	24	282					
cell morphogenesis	GO:0000902	1.394×10^{-2}		26	52	16	282					
cytoskeleton organization	GO:0007010	4.958×10^{-2}		48	52	22	282					

Enrichr output table

Fisher's exact test

GO Biological Process

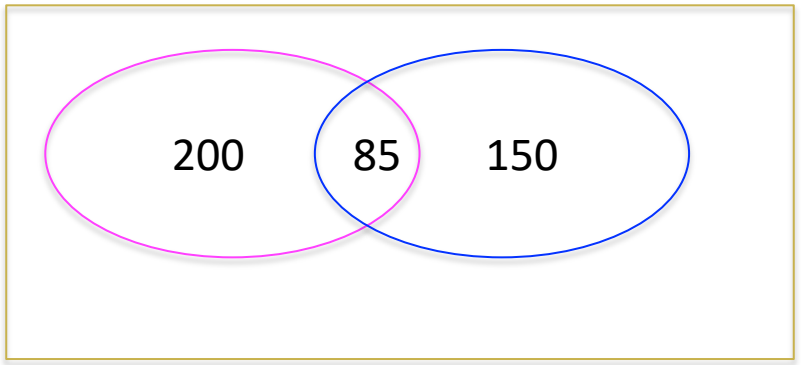
Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Z-score	Combined Score	Genes
extracellular matrix organization (GO:0030198)	85/230	2.1E-50	6.4E-47	4.3E-39	1.3E-35	-1.64651	188.3195	ITGB1;APP;COL16A1;SPARC;COL14A1;
negative regulation of signal transduction (GO:0009968)	58/284	7.2E-20	1.1E-16	2.4E-16	3.6E-13	-1.31194	57.83351	PID1;IRS1;FLT4;PEAR1;GLI3;CYP26B1;
skeletal system development (GO:0001501)	38/147	4.9E-17	4.9E-14	8.3E-14	6.2E-11	-1.47253	55.30609	DLX5;COL12A1;CHRD;AEBP1;PCSK5;PT
regulation of cell migration (GO:0030334)	57/317	7.2E-17	5.5E-14	5.4E-14	5.5E-11	-1.27044	47.21385	ROBO4;SERPINE1;LDB2;FGF1;RND3;CY
collagen fibril organization (GO:0030199)	18/30	2.4E-16	1.5E-13	6.5E-12	3.6E-09	-1.57943	56.77949	LUM;COL14A1;COL11A1;COL12A1;DP
glycosaminoglycan biosynthetic process (GO:0006024)	29/100	9.5E-15	4.8E-12	7.1E-12	3.6E-09	-1.2711	41.04479	CHPF;SDC2;XYLT1;HS2ST1;ACAN;NDST
regulation of angiogenesis (GO:0045765)	38/178	4.1E-14	1.8E-11	1.3E-11	5.4E-09	-1.77078	54.58956	SEMA5A;ITGB1;ECM1;SPARC;SERPINE
positive regulation of cell motility (GO:2000147)	36/180	1.5E-12	5.7E-10	2.1E-10	8E-08	-1.22301	33.29297	LRRC15;SEMA7A;SEMA3C;SEMA3D;TW
protein complex subunit organization (GO:0071822)	18/46	3.6E-12	1.2E-09	1.5E-09	3.6E-07	-1.44324	38.01215	LUM;COL14A1;COL11A1;COL12A1;DP

Pathways (gene-sets)

Overlap:
Numerator -> genes in my gene list and tested pathway

Denominator -> Genes in the original pathway

List of genes in the overlap



PANTHER output

of genes in original pathway

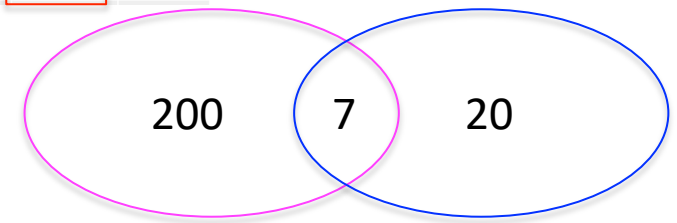
Overlap:# of genes in my gene list and tested pathway

Significance of the enrichment.

Pathway (gene-sets)

Displaying only results for FDR P < 0.05, click here to display all results

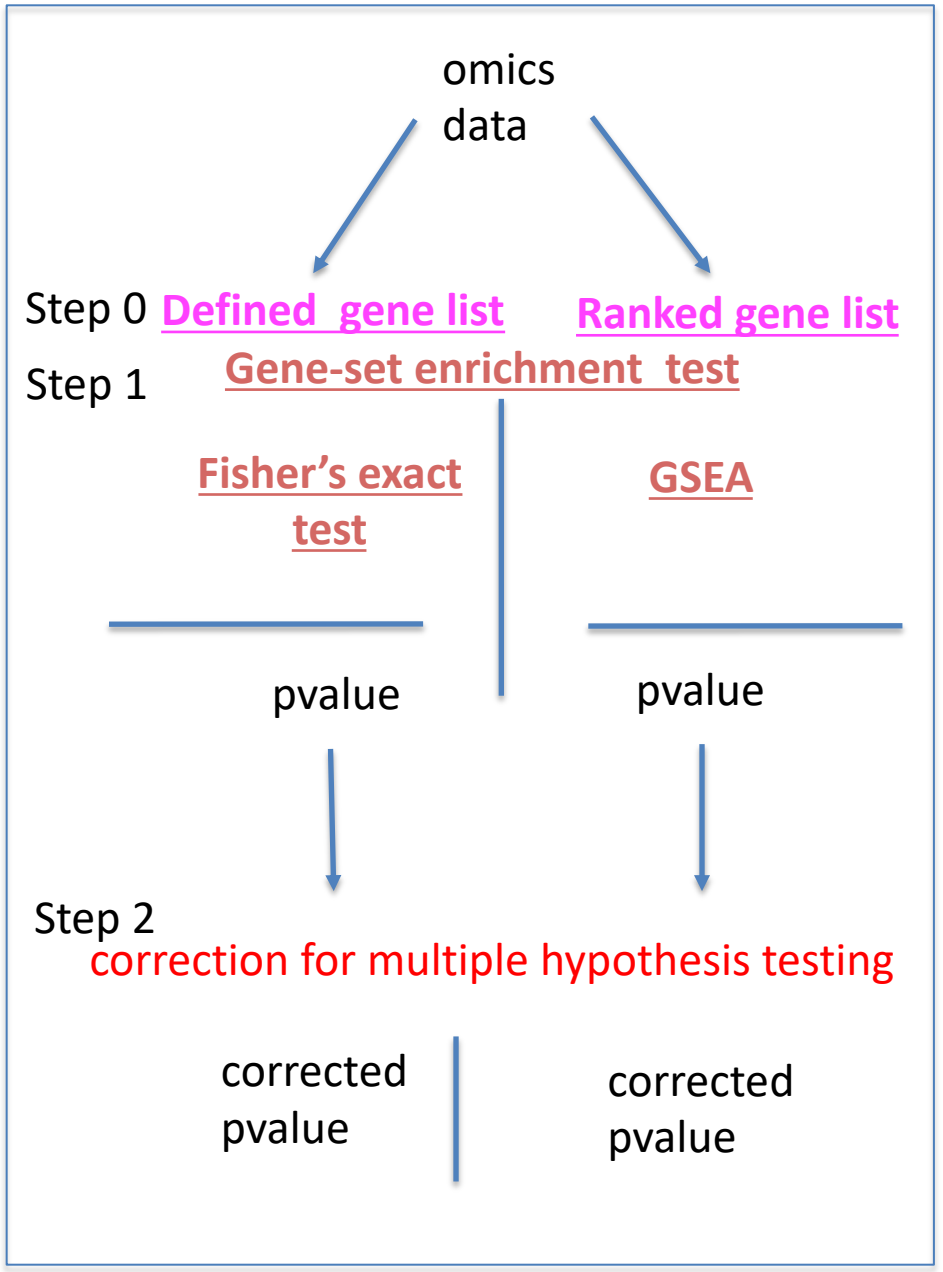
	Homo sapiens (REF)		Client Text Box Input (Hierarchy) NEW!				
PANTHER GO-Slim Biological Process	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
tissue morphogenesis	27	7	1.31	5.33	+	8.09E-04	1.75E-02
regulation of phosphorus metabolic process	250	25	12.16	2.06	+	1.29E-03	2.66E-02
actin filament bundle organization	39	8	1.90	4.22	+	1.31E-03	2.64E-02
regulation of phosphate metabolic process	250	25	12.16	2.06	+	1.29E-03	2.63E-02
regulation of cell communication	359	47	17.46	2.69	+	1.17E-08	1.61E-06
ameboidal-type cell migration	25	8	1.22	6.58	+	1.02E-04	3.17E-03
glycoprotein biosynthetic process	101	13	4.91	2.65	+	2.41E-03	4.33E-02
response to growth factor	75	16	3.65	4.39	+	4.01E-06	1.80E-04
regulation of cell size	28	7	1.36	5.14	+	9.71E-04	2.05E-02
multicellular organism development	609	84	29.61	2.84	+	6.18E-16	2.78E-13
cell-cell signaling	523	47	25.43	1.85	+	1.58E-04	4.37E-03
extracellular matrix organization	69	31	3.36	9.24	+	8.89E-18	1.60E-14
neuron differentiation	224	29	10.89	2.66	+	7.03E-06	2.87E-04
vasculature development	38	13	1.85	7.04	+	3.92E-07	3.20E-05
carbohydrate derivative metabolic process	282	27	13.71	1.97	+	1.54E-03	3.03E-02
cell differentiation	302	38	14.69	2.59	+	6.17E-07	4.26E-05
cellular response to stimulus	1977	140	96.14	1.46	+	1.62E-05	5.83E-04
cell-substrate adhesion	54	10	2.63	3.81	+	6.83E-04	1.51E-02
response to endogenous stimulus	116	16	5.64	2.84	+	4.11E-04	9.84E-03
regulation of Wnt signaling pathway	40	9	1.95	4.63	+	3.69E-04	8.95E-03
regulation of intracellular signal transduction	293	31	14.25	2.18	+	1.44E-04	4.05E-03



Enrichment Analysis using a Ranked Gene List

Outline

- Two types of gene lists (ranked or not)
- Introduction to enrichment analysis
- 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 in ranked gene lists?

- Possible problems with gene list test
  - No “natural” value for the threshold
  - Different results at different threshold settings
  - Possible loss of statistical power due to thresholding
    - No resolution between significant signals with different strengths
    - Weak signals neglected

Example 2: enrichment analysis using a ranked gene list

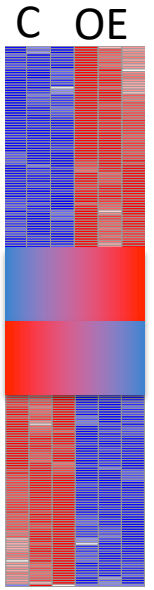
Cells: control
Cells: overexpression of a transcription factor TF

↓

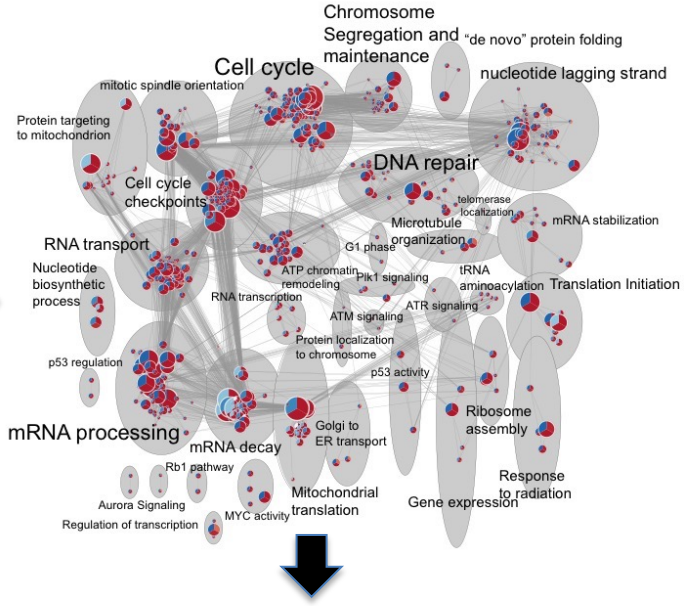
Bulk RNA

↓

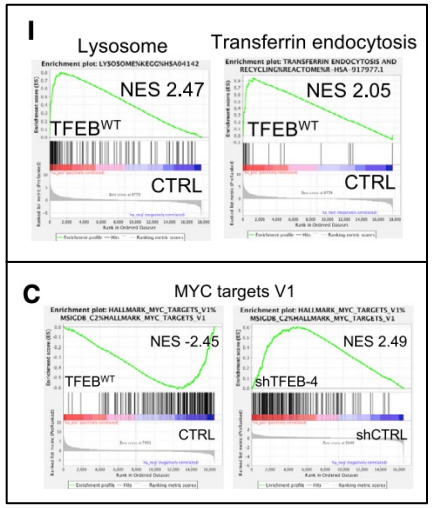
Differential expression analysis



Rank file



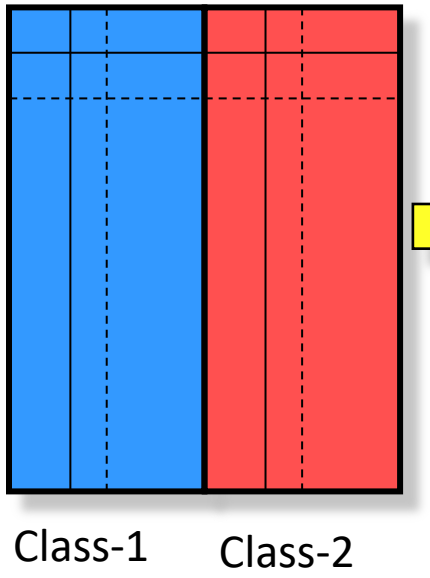
↓



Ref: https://www.cell.com/cell-stem-cell/pdf/S1934-5909(21)00288-5.pdf

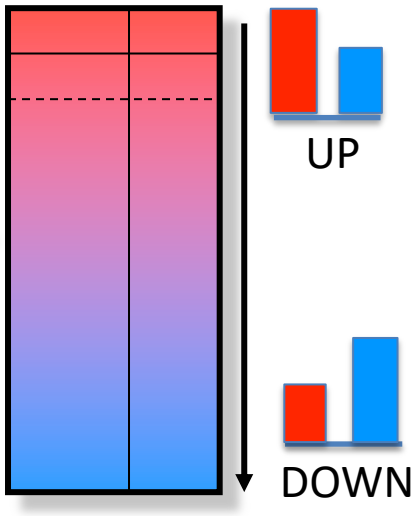
Two-class design : ranked gene list

Expression Matrix



Class-1 Class-2

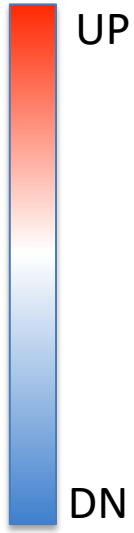
Genes Ranked by Differential Statistic



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



- In their 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 individual genes in this pathway were only down-regulated by a small amount but the addition of all these subtle decreases had a great impact on the pathway.
- They validated their finding experimentally.

http://www.people.vcu.edu/~mreimers/HTDA/Mootha%20-%20GSEA.pdf

Ranked gene list enrichment test

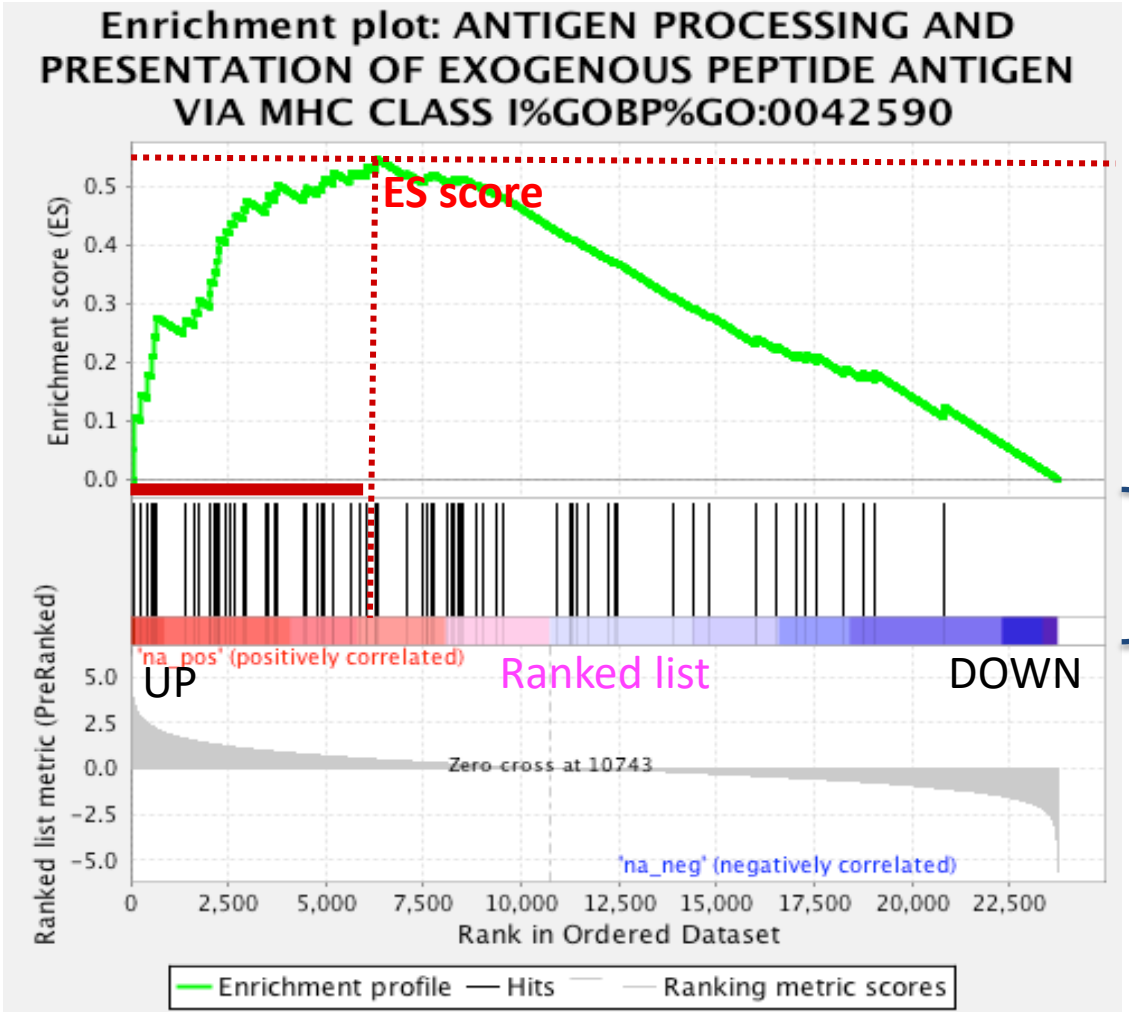
GSEA → modified Kolmogorov Smirnov test (KS test)

https://en.wikipedia.org/wiki/**Andrey_Kolmogorov**#/media/File:Kolm_complexity_lect.jpg

GSEA score calculation

Ranked gene list

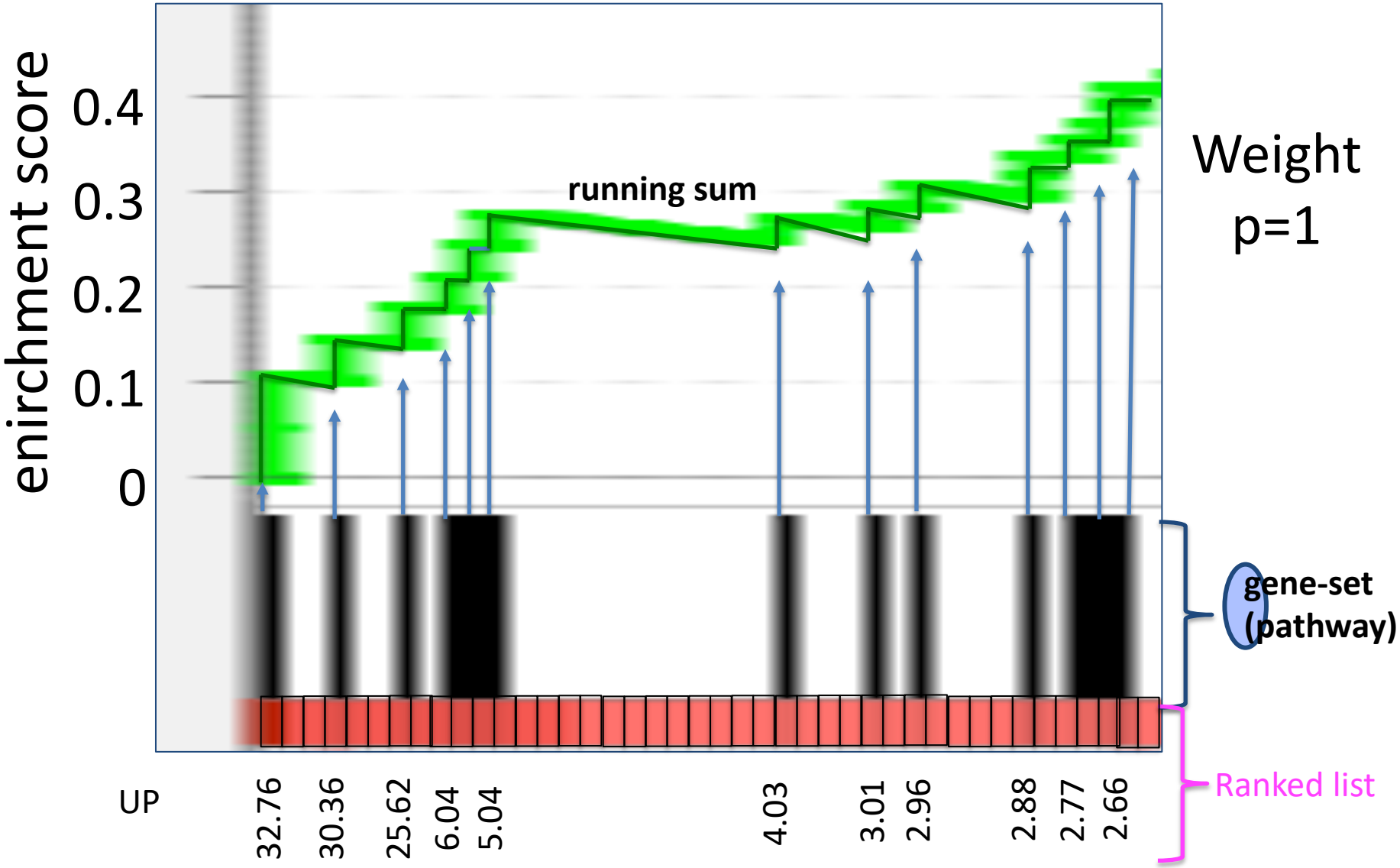
	UP
BGN	32.76
ANTXR1	30.36
FZD1	29.36
COL16A1	28.88
KLF3	1.08
RASEF	0.05
…	…
…	…
ISOC1	0.05
ANO1	0.04
CBWD3	-1.09
GBP4	-15.6
TAP1	-19
PSMB9	-19.7
	DOWN



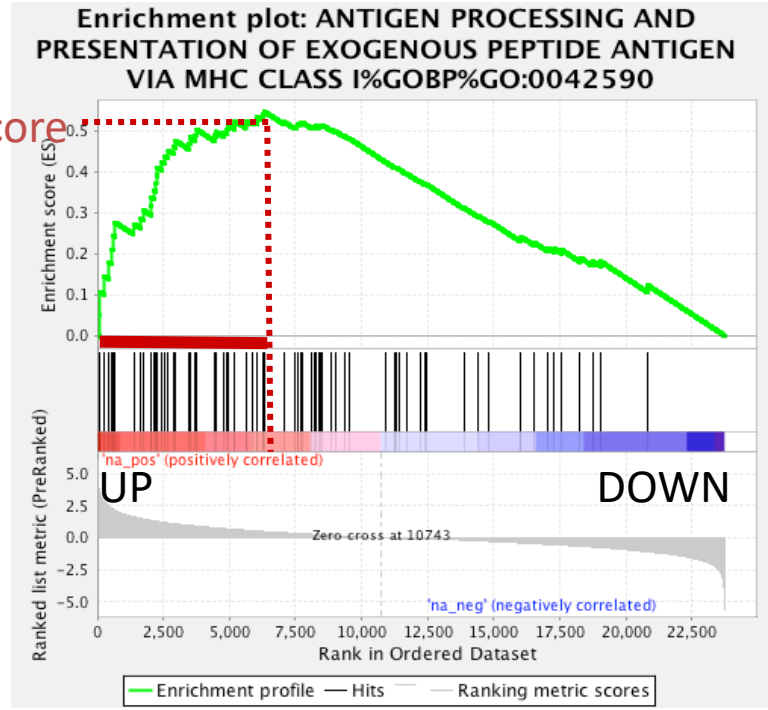
gene-set (pathway)

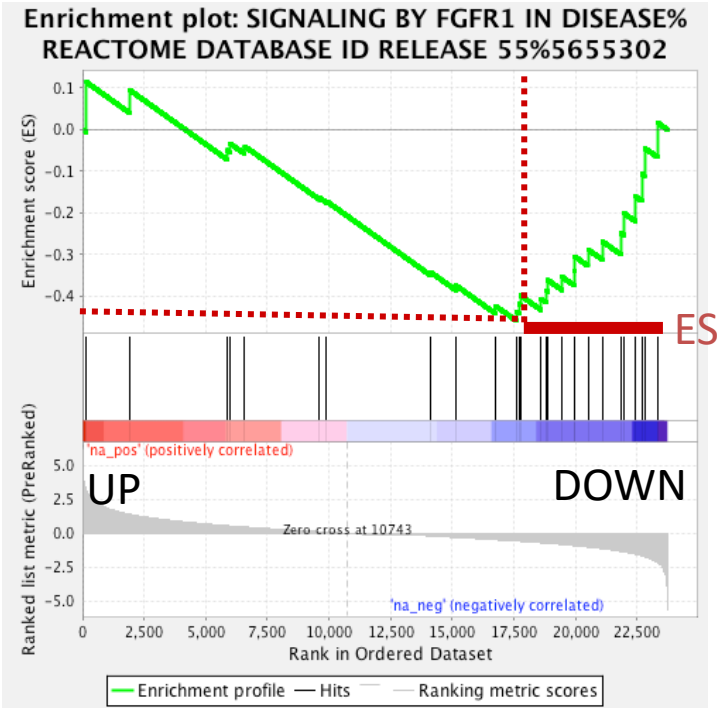
1. Maximum (or minimum) ES score is the final **ES score** for the gene set
2. Can define "leading edge subset" as all those genes ranked as least as high as the enriched set.

GSEA running sum



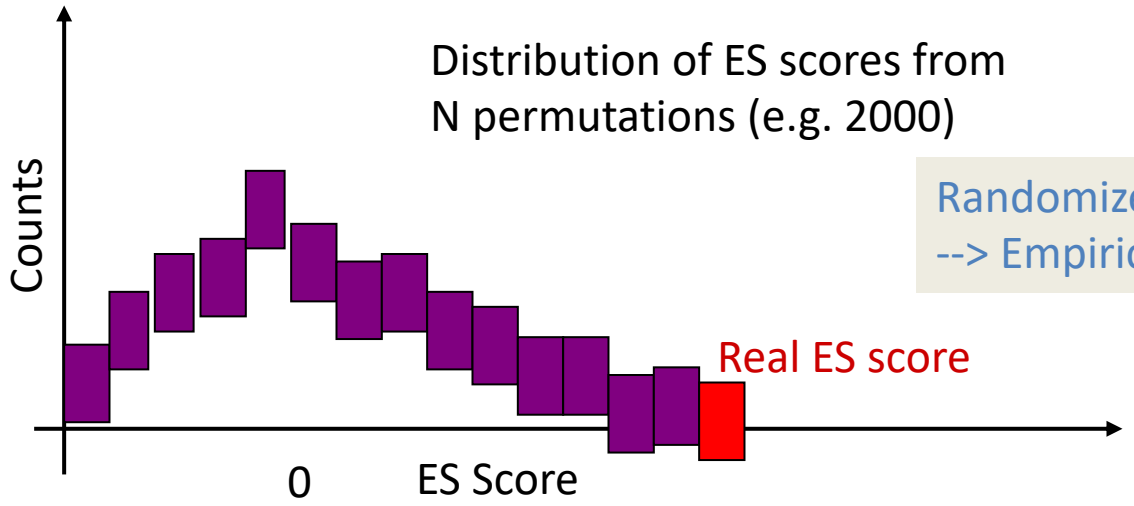
Positive and negative enrichment scores





Going from ES score → P-value → FDR

1. Generate null-hypothesis distribution from randomized data (see permutation settings)
2. Estimate empirical p-value by comparing observed ES score to null-hypothesis distribution from randomized data (for every gene-set)



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

OMICS gene lists: ranked or not ranked?
a few examples

OMICS gene lists: ranked or not ranked?

a few examples

Experimental design: 2 class-design, treated versus control

Starting point:

RNAseq

RNA

↓

Differential expression between treated and control

↓

Ranked list of **all genes** by differential expression score

Single cell RNA seq

RNA

↓

Cell clusters

↓

Differential expression between cluster 1 and cluster 2

↓

Ranked list of **all genes** by differential expression score

Label free proteomics

proteins

> 5,000 proteins

↓

Differential expression between treated and control

↓

Ranked list of **all proteins** by differential expression score

OMICS gene lists: ranked or not ranked?

a few examples, cont.

Start point is: DNA

Looking for somatic mutations or CNV

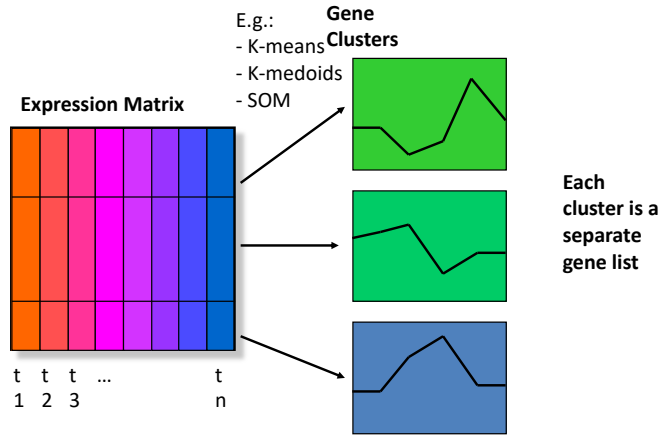
↓

Variant calling

gene list

(eg list of frequently mutated genes)

RNA: Time course or cluster analysis



ATAC-seq

Chip-seq

↘ ↙

Peaks regions
(BED FILE)

↓

Need to associate peak regions with genes

↓

Gene list of associated with peaks of interest

Many available enrichment analysis tools

web-based

Cytoscape app

Standalone

R package

Typical output of an enrichment analysis is:

Pathway name	Number of overlapping genes	Number of genes in pathway	P-value	Adjusted p-value
…	….	…	…	….

Typical output

gene-set name (pathway)

number of overlapping genes

… corrected for gene-set size

p-value

… corrected for multiple hypothesis

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 PEPTIDE LOADING OF CLASS I MHC%REACTOME%REACT_7	24	1.76	0.0041	0.0505599
ABORTIVE ELONGATION OF HIV-1 TRANSCRIPT IN THE ABSENCE OF TAT%REACTOME%REACT_6261.3	23	1.75	0	0.0529242
DNA DAMAGE RESPONSE, SIGNAL TRANSDUCTION BY P[illegible] CLASS MEDIATOR RESULTING IN CELL CYCLE ARREST%	67	1.75	0	0.052886
REGULATION OF MACROPHAGE ACTIVATION%GO%GO:00[illegible]	11	1.75	0.003	0.0534709
PROTEIN FOLDING%REACTOME%REACT_16952.2	52	1.75	0.002	0.0537717
ENDOPLASMIC RETICULUM UNFOLDED PROTEIN RESPONSE%GO%GO:0030968	73	1.75	0	0.0546052
PROTEIN EXPORT%KEGG%HSA03060	24	1.75	9.75E-04	0.0548699
TRANSCRIPTION INITIATION FROM RNA POLYMERASE I PROMOTER%GO%GO:0006367	64	1.75	0.001	0.0545783
S PHASE%REACTOME%REACT_899.4	110	1.75	0	0.0546003
PROTEASOMAL PROTEIN CATABOLIC PROCESS%GO%GO:001[illegible]	163	1.75	0	0.0550066
ATP-DEPENDENT RNA HELICASE ACTIVITY%GO%GO:0004004	20	1.74	0.0059	0.0556722
ACID-AMINO ACID LIGASE ACTIVITY%GO%GO:0016881	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%BIOCARTA_IL4_PATHWAY	11	1.74	0.0082	0.0581508
ASSOCIATION OF TRIC CCT WITH TARGET PROTEINS DURING BIOSYNTHESIS%REACTOME%REACT_16907.2	28	1.74	0.0039	0.0581298
UBIQUITIN-DEPENDENT DEGRADATION OF CYCLIN [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_17[illegible]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
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABULAR FORMAT

NETWORK VISUALIZATION

How to choose a tool?

- Does it cover your model organism?
- Is there a good choice of gene-sets (pathway database)
- Are the pathway databases up to date?
- Which statistics (for gene list or ranked gene list)?
- Is the description of statistics clear enough ?
- Do you like the output style?
- Can you connect it with network visualization tools like Cytoscape?

Defined gene list (Fisher's exact test)

	g:Profiler	PANTHER	biNGO	Cluego
Updated database	yes	yes	no? *1	yes
Choice of database (more than 1)	yes	yes	no (GO) *1	yes
Do we test database individually or together	together	individually	individually	together
Multiple model organisms?	yes	yes	yes	yes
Possibility to upload your own custom database	yes	no?	yes	no?
Statistics: possibility to use the Fisher's exact test (ORA) (thresholded gene list)	yes	yes	yes	yes
Multiple hypothesis correction; possibility to use B-H FDR	yes	yes	yes	yes
Possibility to upload reference genes (background)	yes	yes	yes	yes
Website (Web) or Cytoscape App (App)	Web	Web	App	App
Possibility to visualize with Cytoscape EnrichmentMap	YES	no	YES	Cytoscape

*1: can still be used with custom database ;

Notes

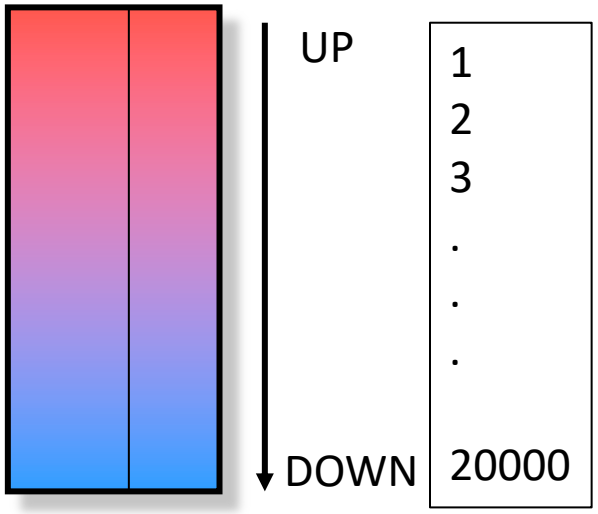
- We usually test **over-enrichment** of “black”. To test for ***under-enrichment*** of “black”, test for *over-enrichment* of “red”.
- **Fisher’s Exact Test** is often called the **hypergeometric test**
- **Other enrichment tests** for defined gene lists (not covered in this lecture):
  - Approximation of the Fisher’s Exact Test (Monte Carlo simulation)
  - Binomial test
  - Chi-squared test

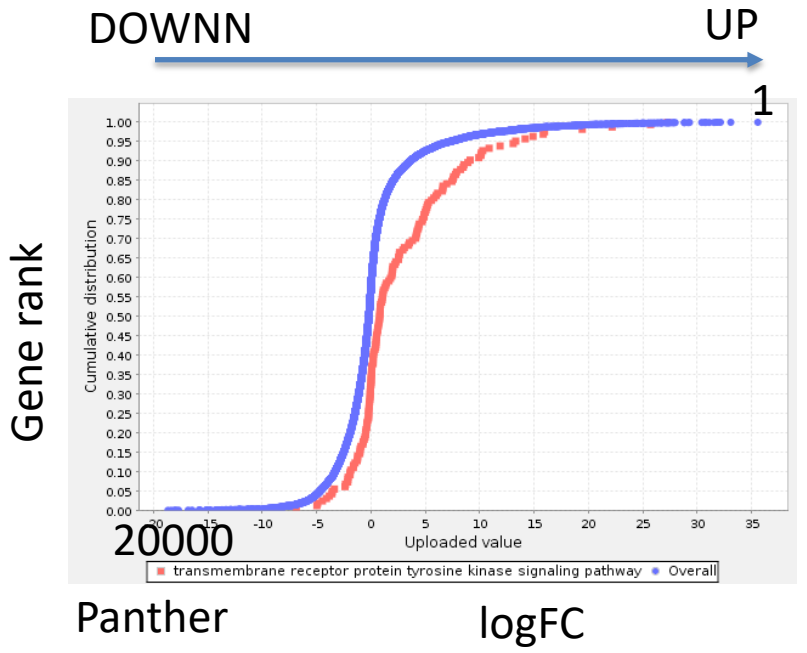
Ranked list

	GSEA	PANTHER
Rank test	Modified KS test	Wilcoxon Rank Sum test
Correction for multiple hypothesis testing	yes	yes
Choice of gene-sets + able to custom pathway database , can therefore be use for different model organisms	yes	no
Possibility to visualize results with Cytoscape enrichment map	yes	no

Other enrichment tests for a ranked gene list

Wilcoxon ranksum test





Outline of theory component

- Fisher's exact test (or binomial) for calculating enrichment P-values for defined gene lists
- GSEA, wilcoxon rank sum test for computing enrichment P-values for ranked gene lists

Recipe for defined gene list enrichment test

- **Step 1:** Define your gene list and your background list,
- **Step 2:** Select your gene sets (pathways) to test for enrichment,
- **Step 3:** Run enrichment tests using the Fisher's exact test and correct for multiple testing if you test more than one gene set (pathway)
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Recipe for ranked list enrichment test

- **Step 1:** Rank your genes,
- **Step 2:** Select your gene sets (pathways) to test for enrichment,
- **Step 3:** Run enrichment tests and correct for multiple testing, if necessary,
- **Step 4:** Interpret your enrichments
- **Step 5:** Publish! ;)

Advanced topics
(not covered in this lecture)

- Issues with tests: correlation between gene-sets, dependency of genes.
- Other types of tools: topology aware.
- Modern tools are starting to include some network visualization.

Go to: Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap

https://www.nature.com/articles/s41596-018-0103-9

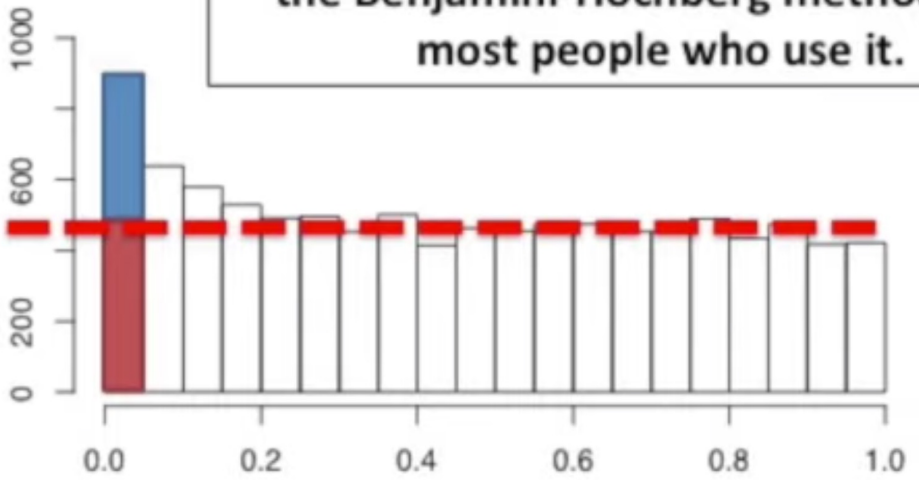
Tips

- Be precise at each step of your analysis
- Try to answer one biological question at a time

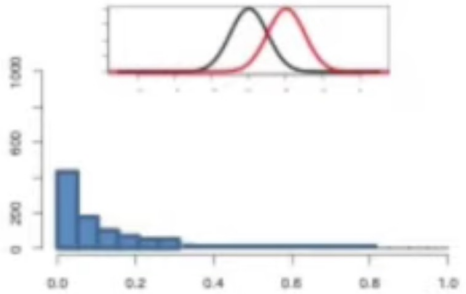
How to win the p-value lottery, part 2

Keep the gene list the same, evaluate different gene-sets(pathways)

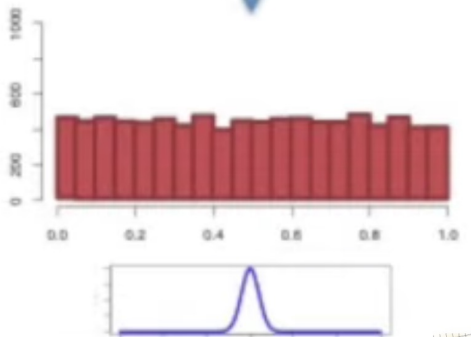
If you can understand these concepts, then you understand more about FDR and the Benjamini-Hochberg method than most people who use it.



P value



p-value distributions



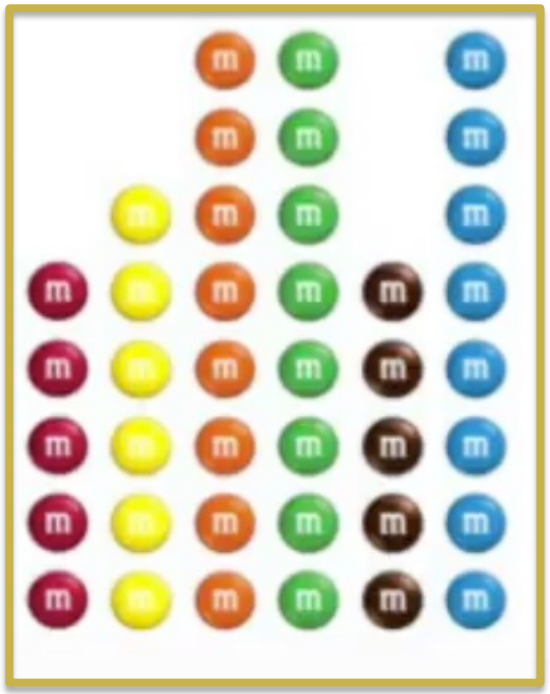
https://www.youtube.com/watch?v=K8LQSvtjcEo

Do you need to learn more about Fisher's exact test?

VIDEO the M&M's examples:
https://www.youtube.com/watch?v=udyAvvaMjfM

StatQuest with Josh Starmer

gene sets



gene list

I'm going to use the histogram of the "ideal" bag of m&m's, based on proportions I got off the internet, and my "sample", my handful of m&m's, to determine if my bag is special

Background

And
Pathway Commons Guide:
https://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR
Ontario Institute for Cancer Research

Ontario Genomics

GenomeCanada