bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by

McGill UNIVERSITY


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

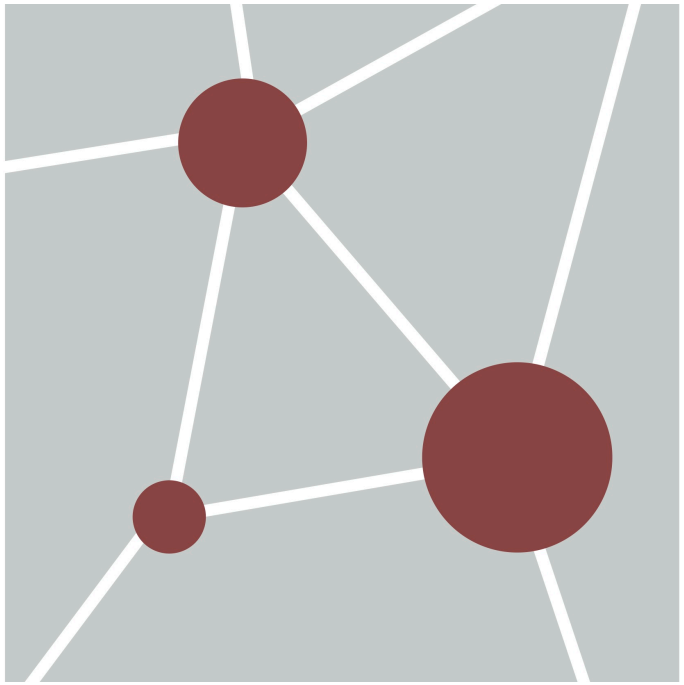
Finding over-represented pathways in gene lists

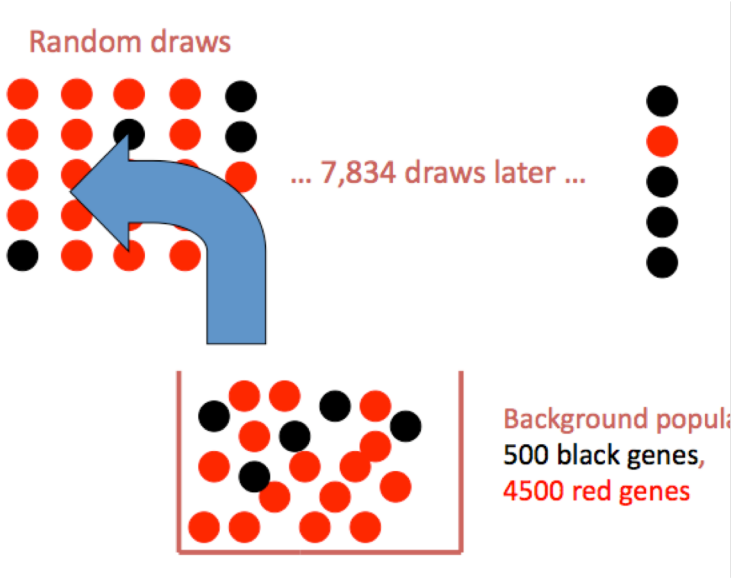
Veronique Voisin

Pathway and Network Analysis of –omics Data

July 27-29, 2020

bioinformatics.ca

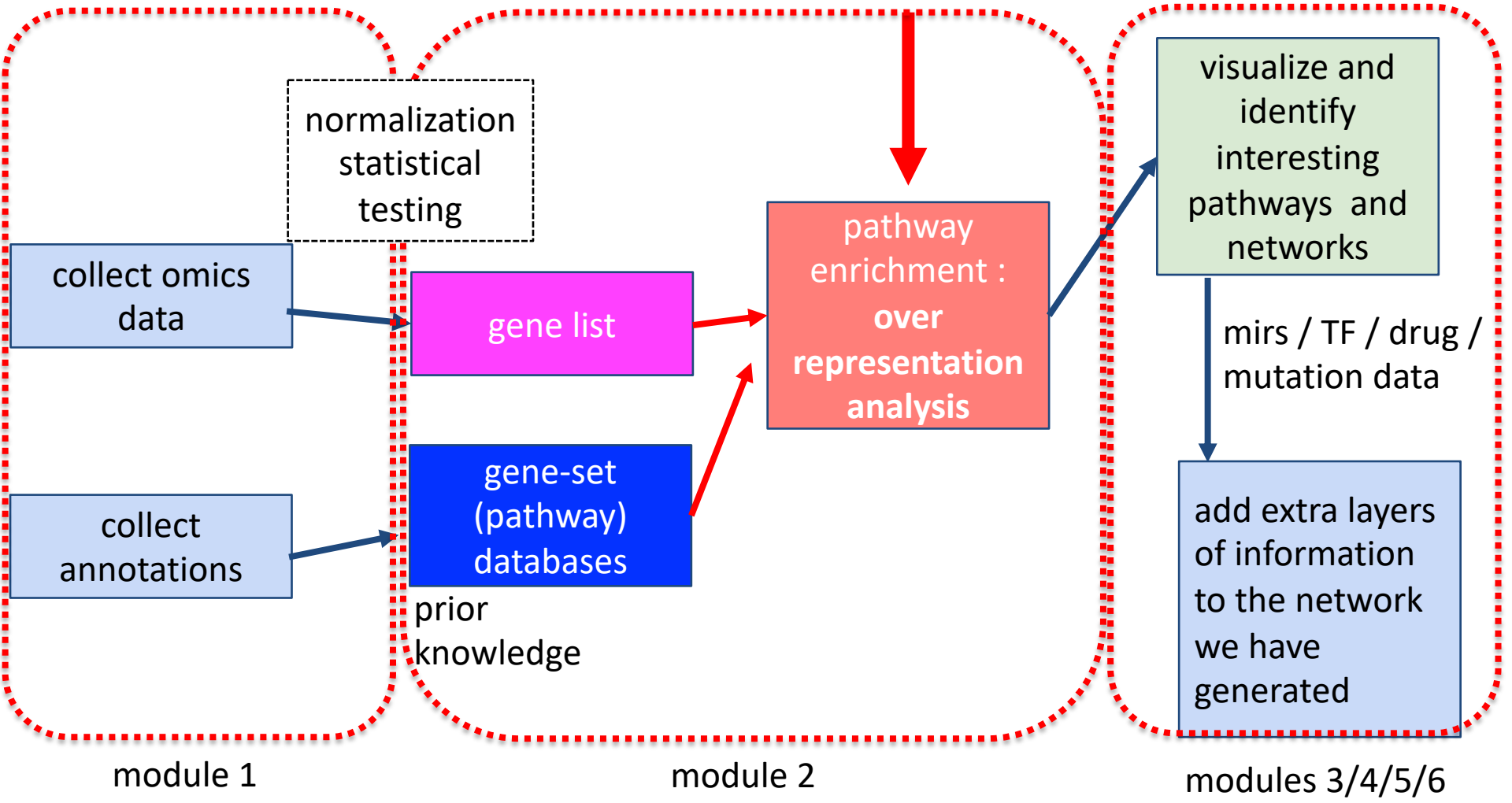




Learning Objectives

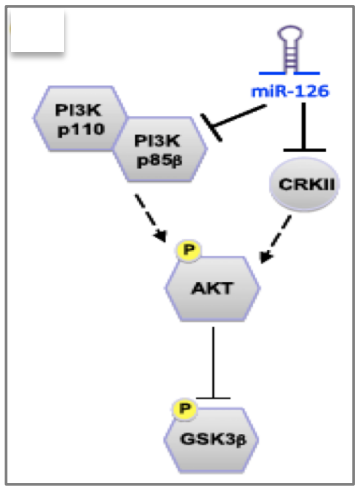
- Be able to understand the differences between a **defined gene list** and a **ranked gene list** and which enrichment test to apply.
- Be able to understand the **result of an enrichment test** and how to interpret it
- Be able to understand the concept of **pvalue** and **corrected pvalue (FDR)**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

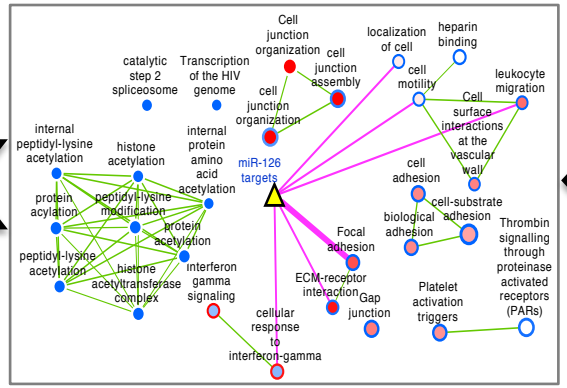
Analysis workflow



pathway analysis workflow…rewind

"In HSC/early progenitors, miR-126 regulates multiple targets within the PI3K/AKT/GSK3β pathway, attenuating signal transduction in response to extrinsic signals. "





enrichment analysis

RNA OR DNA EXTRACTION

↓

HIGH THROUGPUT SEQUENCING

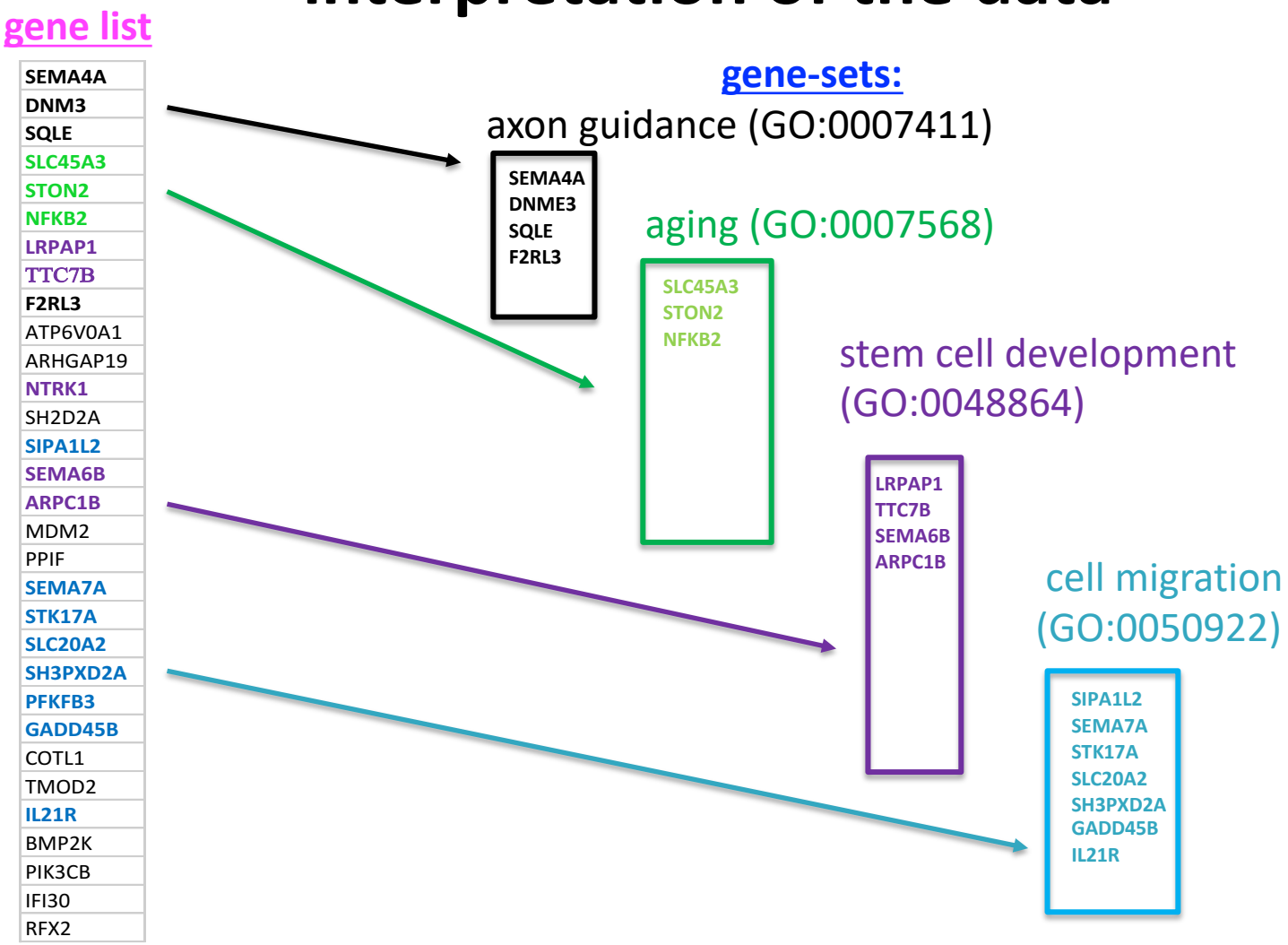
↓

GENE LIST

↓

Focus of this module

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



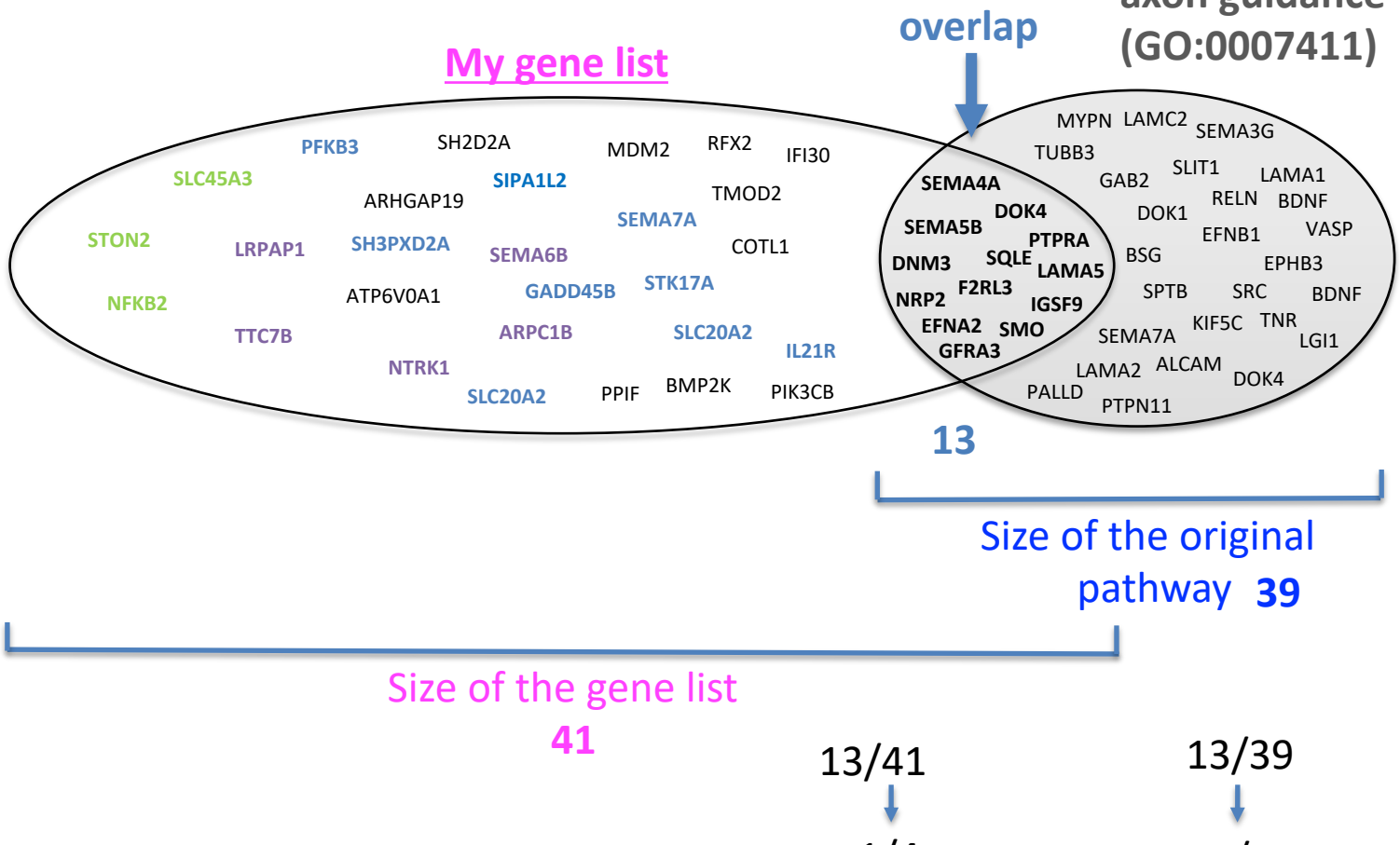
Gene set enrichment analysis calculates the overlap between our gene list and a pathway

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

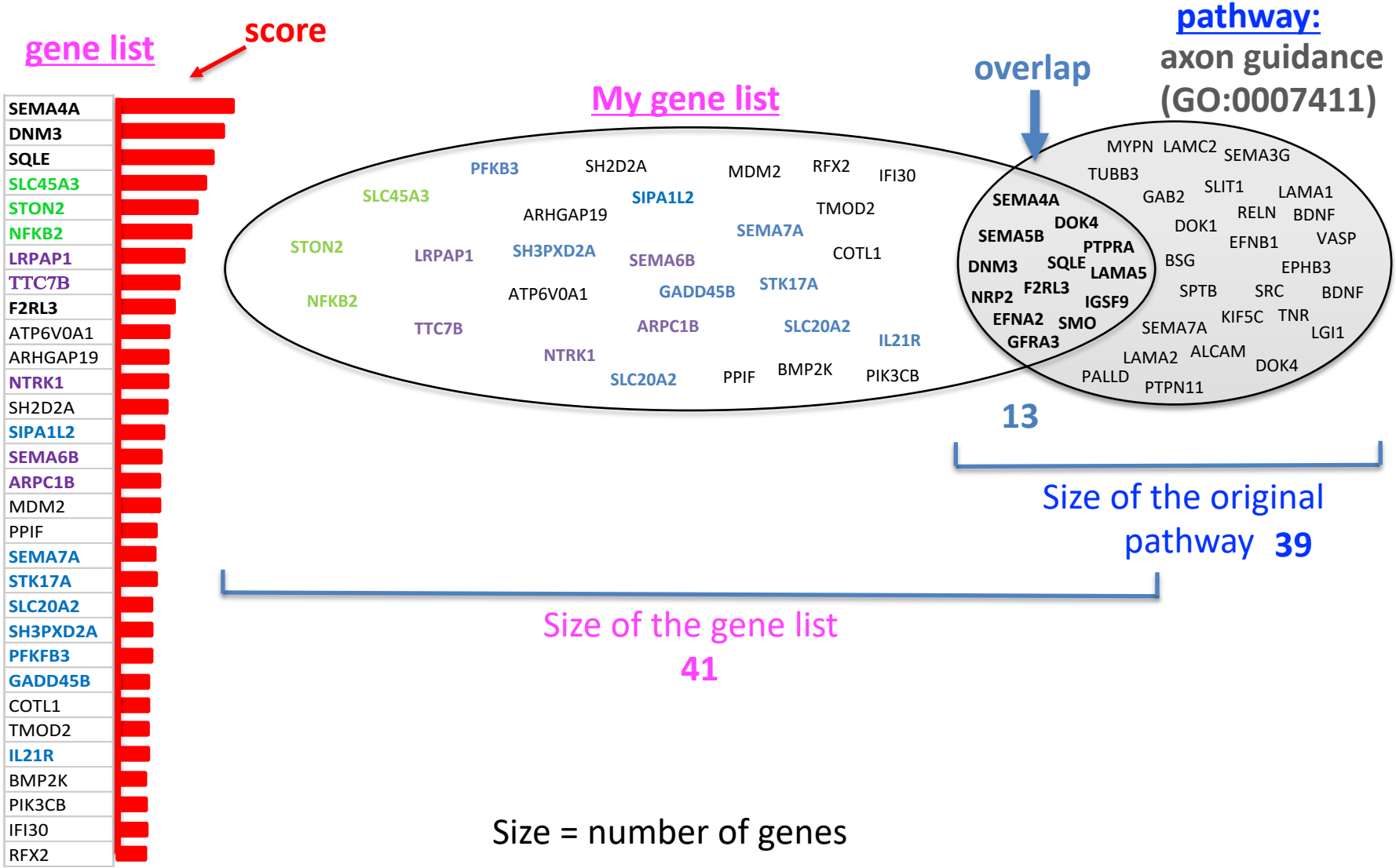


13/41 → 1/4

13/39 → 1/4

Size = number of genes

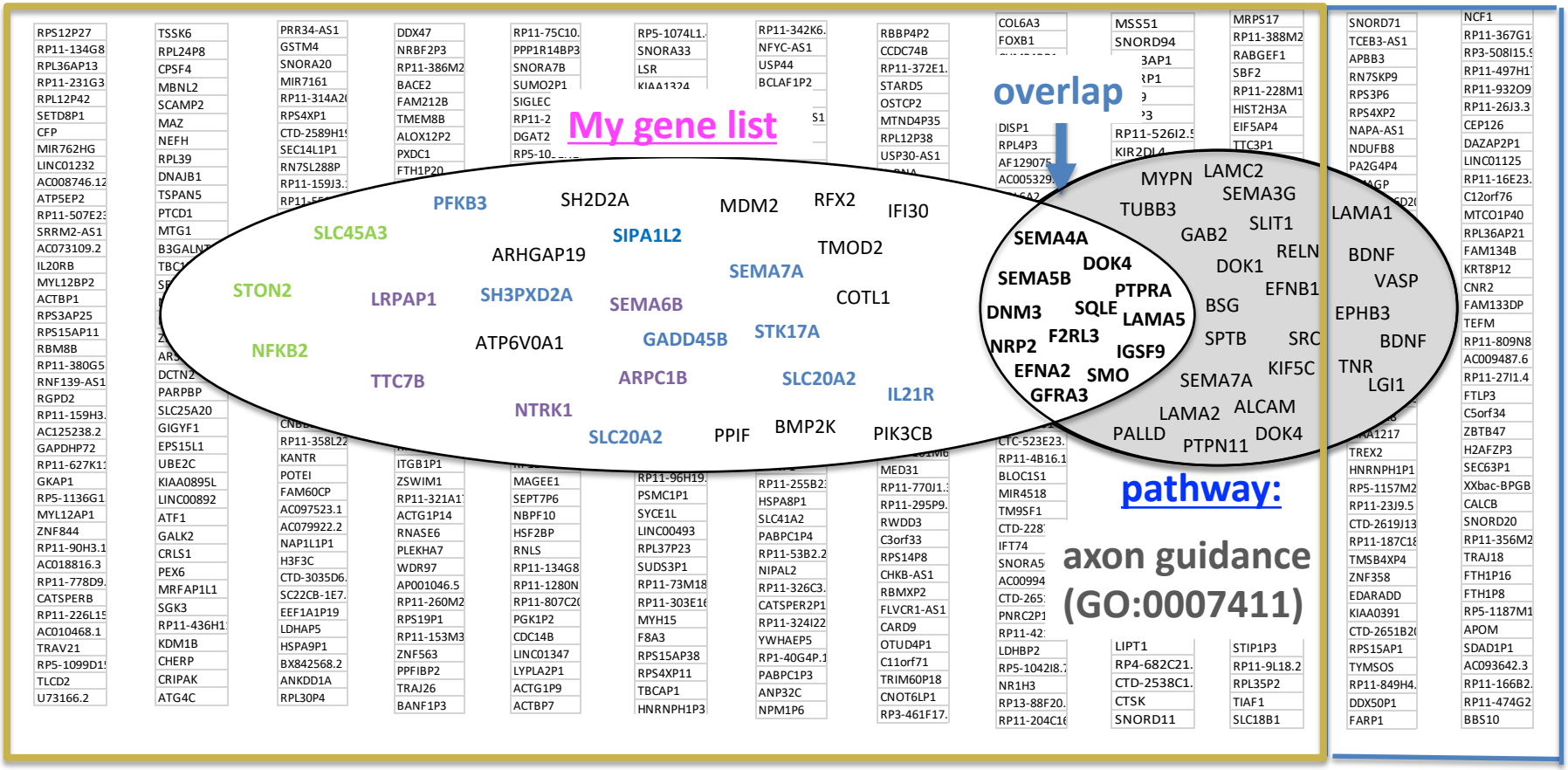
Can we add a score associated with the genes when calculating the enrichment score?



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**

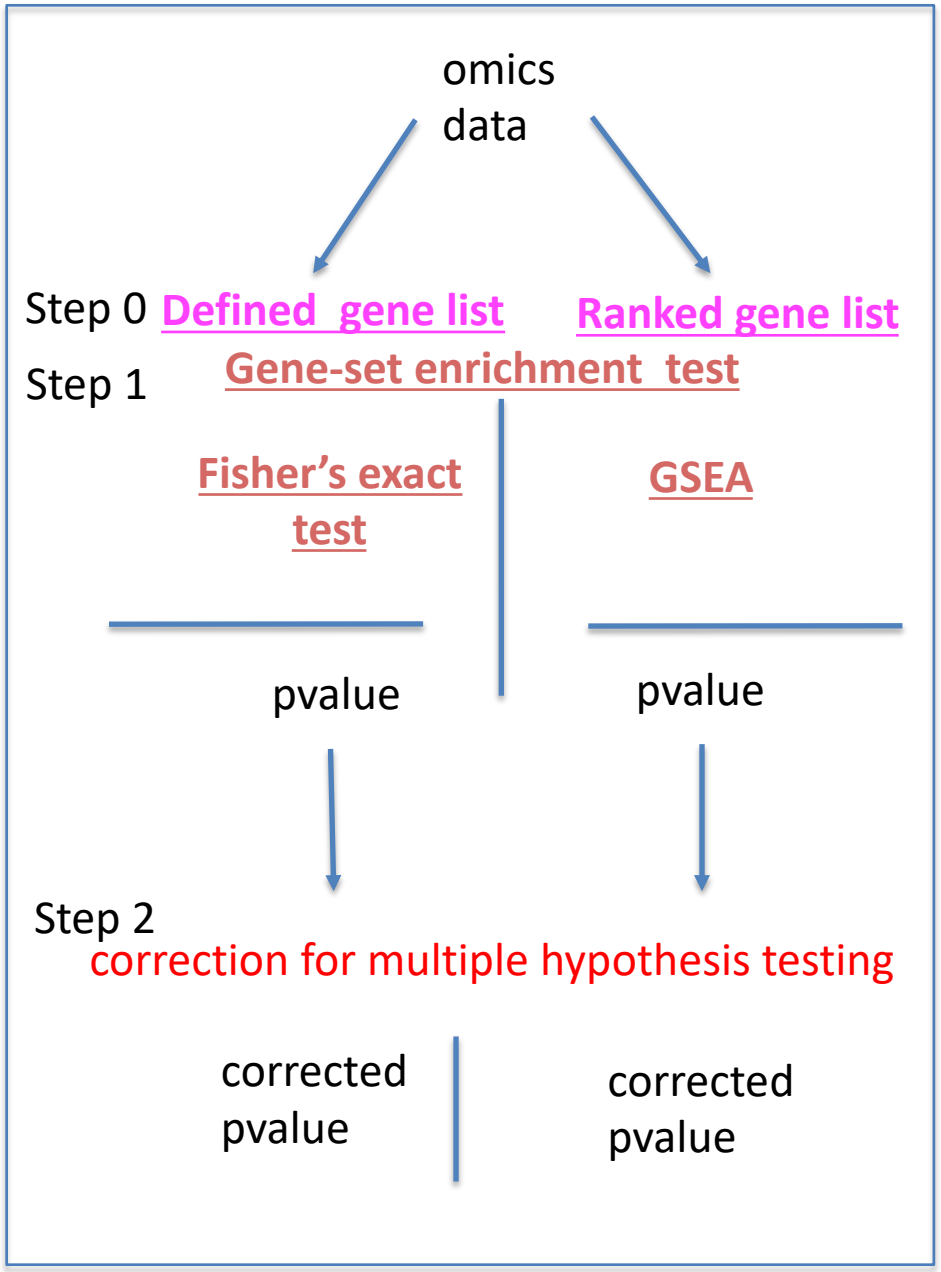
How many genes could have been captured in your experiment?

We are testing many pathways at the same time

correction for multiple hypothesis testing

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)

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

OMICS gene lists: ranked or not ranked?

a few examples

Experimental design: 2 class-design, treated versus control

Starting point:

RNA

RNAseq

↓

Reads are aligned to ref genome

↓

Normalized counts per gene

↓

Ranked list of all genes by differential expression

RNA

Single cell data

↓

Cell clusters

↓

Differential expression between cluster list

↓

Ranked list of all genes by differential expression

RNA

Arrays: Affymetrix or Illumina bead chip

↓

Normalized Intensity per gene/probe

↓

Ranked list of all genes by differential expression

proteins

Label free proteomics > 5,000 proteins

↓

Normalized Intensity per protein

↓

Ranked list of all proteins by differential expression

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)

Start point is: metabolites

Raw spectra

↓

Peak calling

↓

Compound annotation

←

Compound List + gene list (if available)

→ mapped to KEGG metabolic pathways

ATAC-seq

Chip-seq histone

Chip-seq (transcription factors)

↘ ↙

Peaks regions (BED FILE)

↓

Need to associate peak regions with genes

↓

Gene list of associated with peaks of interest

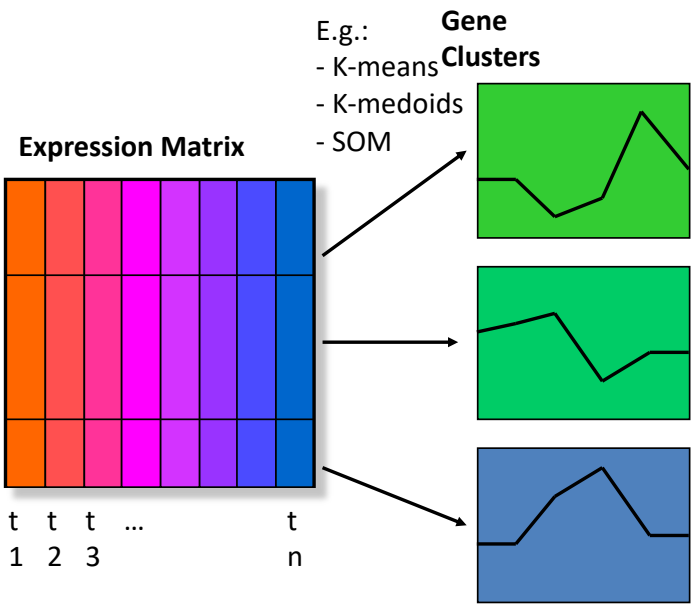
Is there an expected association between chromatin structure and gene expression

↓ If yes and RNAseq data are available

Gene list = Peak associated with genes that are differentially expressed.

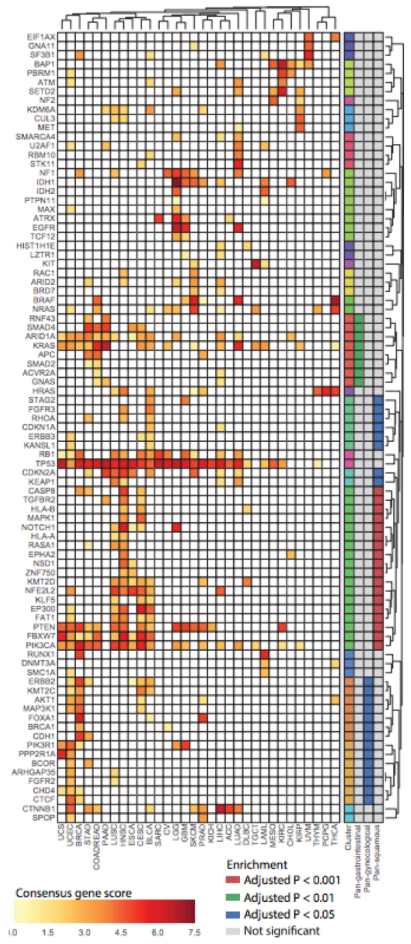
Example of defined gene lists

RNA: Time course or cluster analysis



Each cluster is a separate gene list

DNA: Gene list of frequently mutated genes



Comprehensive Characterization of Cancer Driver Genes and Mutations

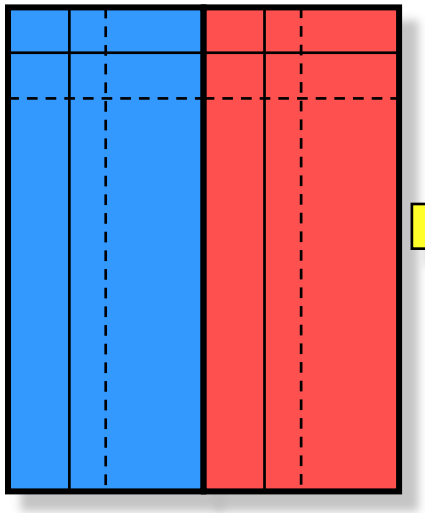
Bailey et al. Cell 2018.
PMID: 30096302

They reported a PanCancer and PanSoftware analysis spanning **9,423 tumor exomes** and using 26 computational tools to catalog driver genes and mutations.

They identified **299 driver genes** with implications regarding their anatomical sites and cancer/cell types

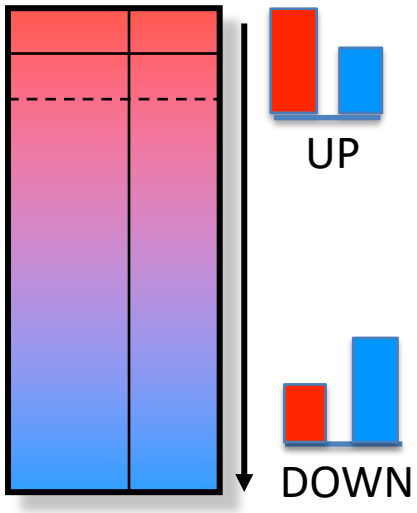
Two-class design : ranked gene list

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)

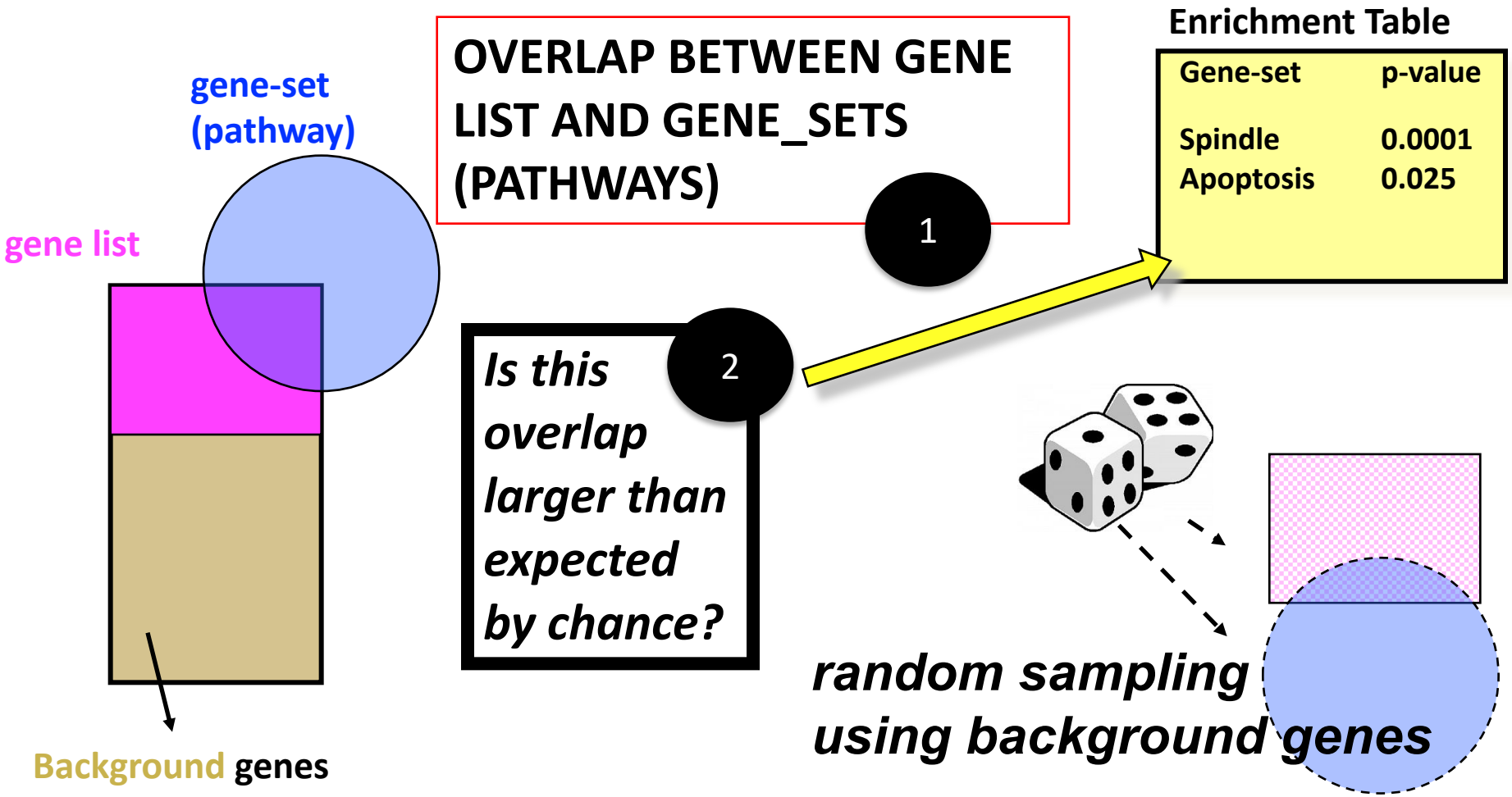
	LogFC	Pvalue	score
BGN	+1	1.73E-33	32.76
ANTXR1	+1	4.39E-31	30.36
FZD1	+1	4.41E-30	29.36
COL16A1	+1	1.33E-29	28.88
KLF3	+1	8.32E-02	1.08
RASEF	+1	9.01E-01	0.05
ISOC1	+1	9.01E-01	0.05
ANO1	+1	9.01E-01	0.04
CBWD3	-1	8.18E-02	-1.09
GBP4	-1	2.45E-16	-15.61
TAP1	-1	1.04E-19	-18.98
PSMB9	-1	1.84E-20	-19.73

Gene list enrichment test

Gene list enrichment analysis

- Given:
  1. **Gene list**: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
  2. **Gene sets (pathways)** or annotations: e.g. The Gene Ontology, transcription factor binding sites in promoter
- Question: ***Are any of the gene sets (pathways) surprisingly enriched in the gene list?***
- Details:
  - Where do the **gene lists** come from?
  - How to assess "surprisingly" (statistics)
  - How to correct for repeating the tests

How do simple enrichment tests work?

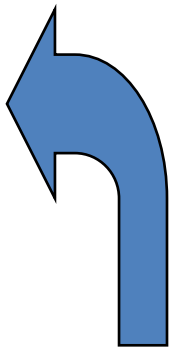


Empirical pval =(#obs_overlap > random_overlap) +1) / (number of tests + 1)

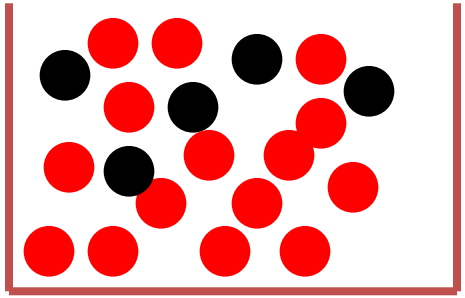
The Fisher's exact test

Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42



Null hypothesis: List is a random sample from population
Alternative hypothesis: More black genes than expected in my list



Background population:
500 black genes,
4500 red genes

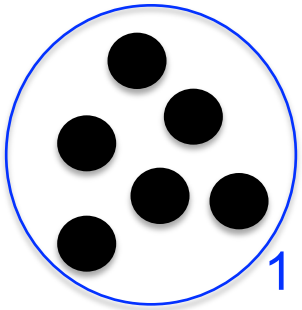
The Fisher's exact test

a.k.a., hypergeometric test

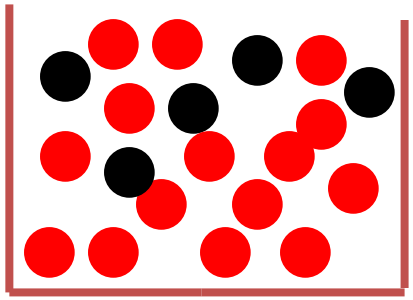
Gene list

- RRP6
- MRD1
- RRP7
- RRP43
- RRP42

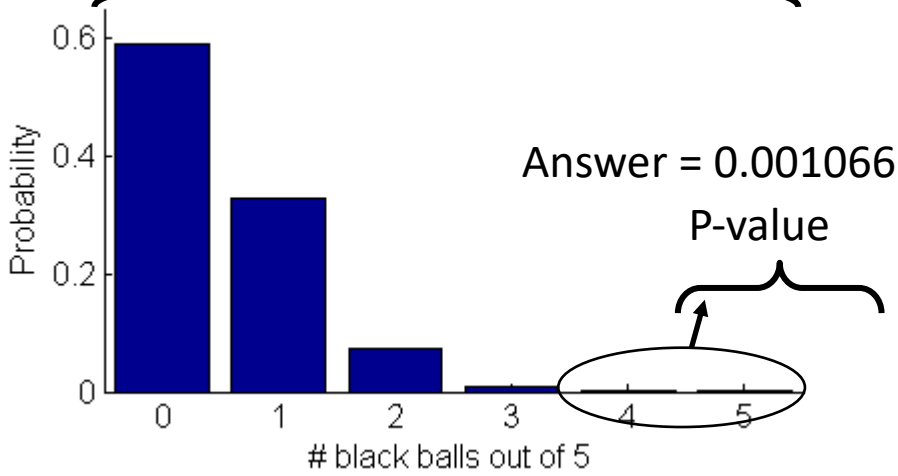
Background population:
500 black genes,
4500 red genes



1 gene-set/pathway

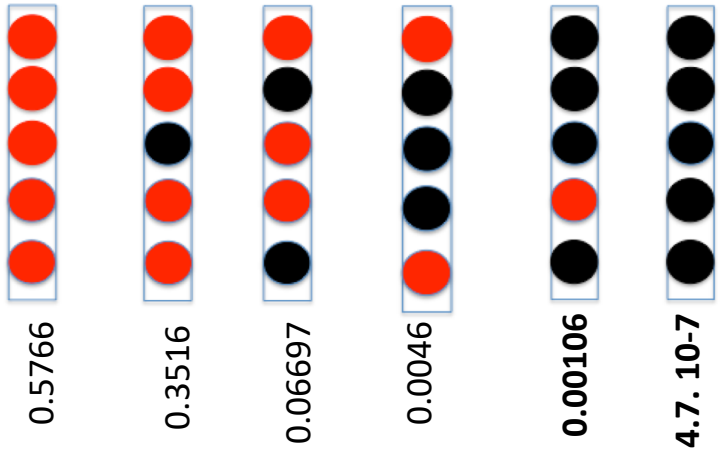


Null distribution



Answer = 0.001066662

P-value



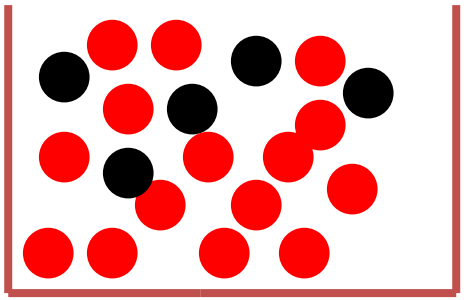
2x2 contingency table for Fisher's Exact Test

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

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



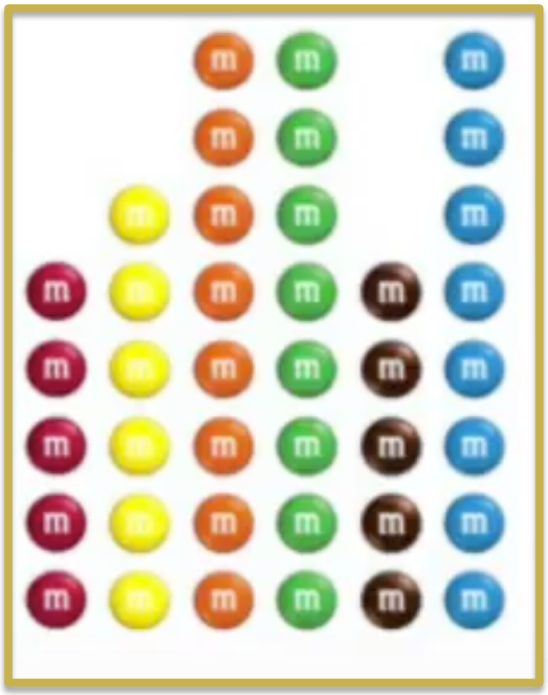
Background population:
500 black genes,
4500 red genes

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/

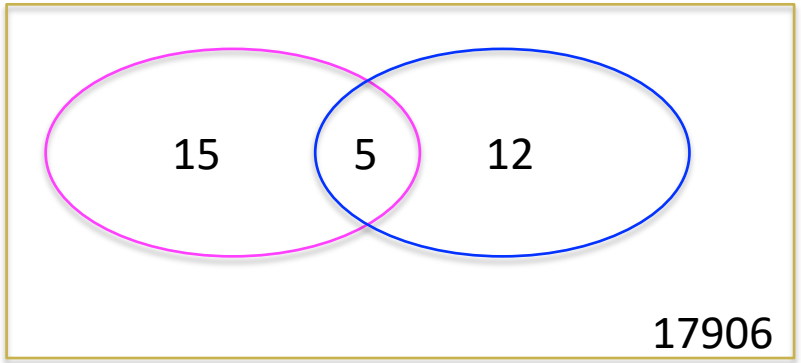
Important points

- We usually test over-enrichment of “black”. To test for *under-enrichment* of “black”, test for *over-enrichment* of “red”.
- Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.
- To test for enrichment of more than one independent types of annotation (red vs black and circle vs square), apply Fisher’s exact test separately for each type.

g:Profiler

GO:BP		stats					
Term name	Term ID	padj	−log10(padj) (0 … ≤16)	T	Q	TnQ	U ↑
pulmonary valve morphogenesis	GO:0003184	1.034×10^{-8}		17	20	5	17906
pulmonary valve development	GO:0003177	3.392×10^{-8}		21	20	5	17906
regulation of myeloid leukocyte differentiation	GO:0002761	6.876×10^{-8}		122	20	7	17906
regulation of osteoclast differentiation	GO:0045670	1.353×10^{-7}		67	20	6	17906

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



2x2 contingency table

	In gene list	Not in gene list
In pathway	5	12
Not in pathway	15	17894
	20	17906

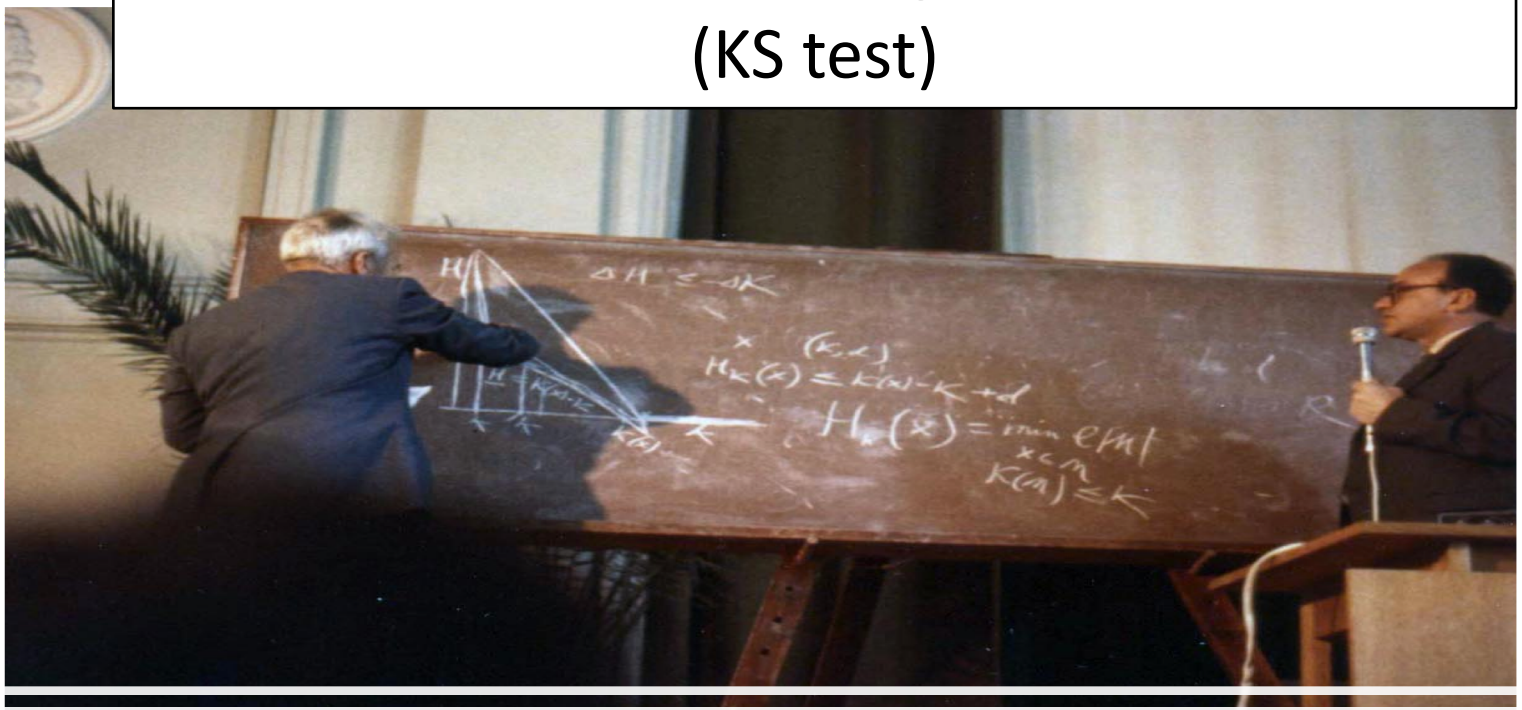
Other enrichment tests for defined gene lists (not covered in this lecture)

Note: Fisher's Exact Test is often called the hypergeometric test

- Approximation of the Fisher's Exact Test (Monte Carlo simulation)
- Binomial test
- Chi-squared test

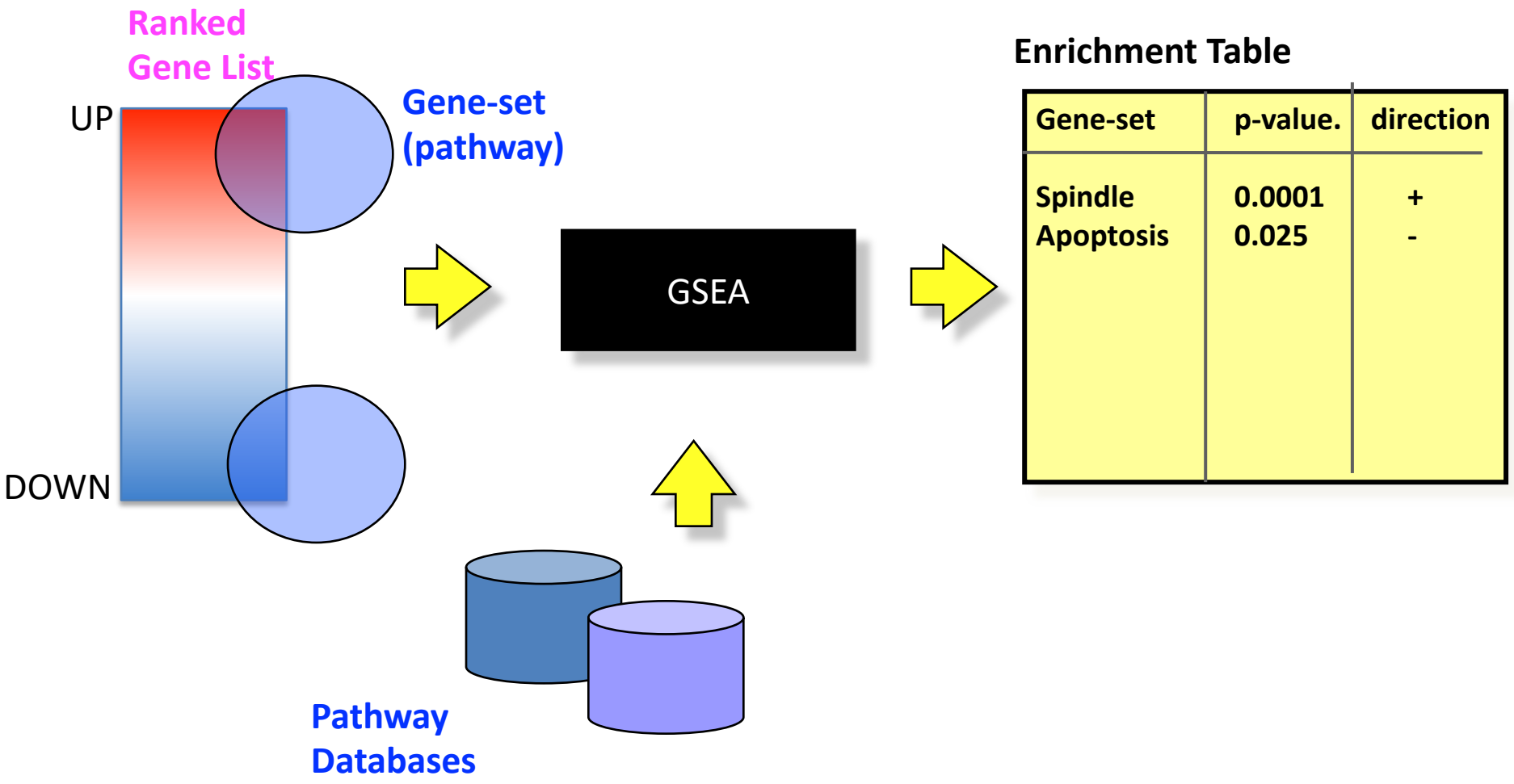
Ranked gene list enrichment test

GSEA → modified Kolmogorov Smirnov test (KS test)



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

Example of a ranked list enrichment test



- 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.

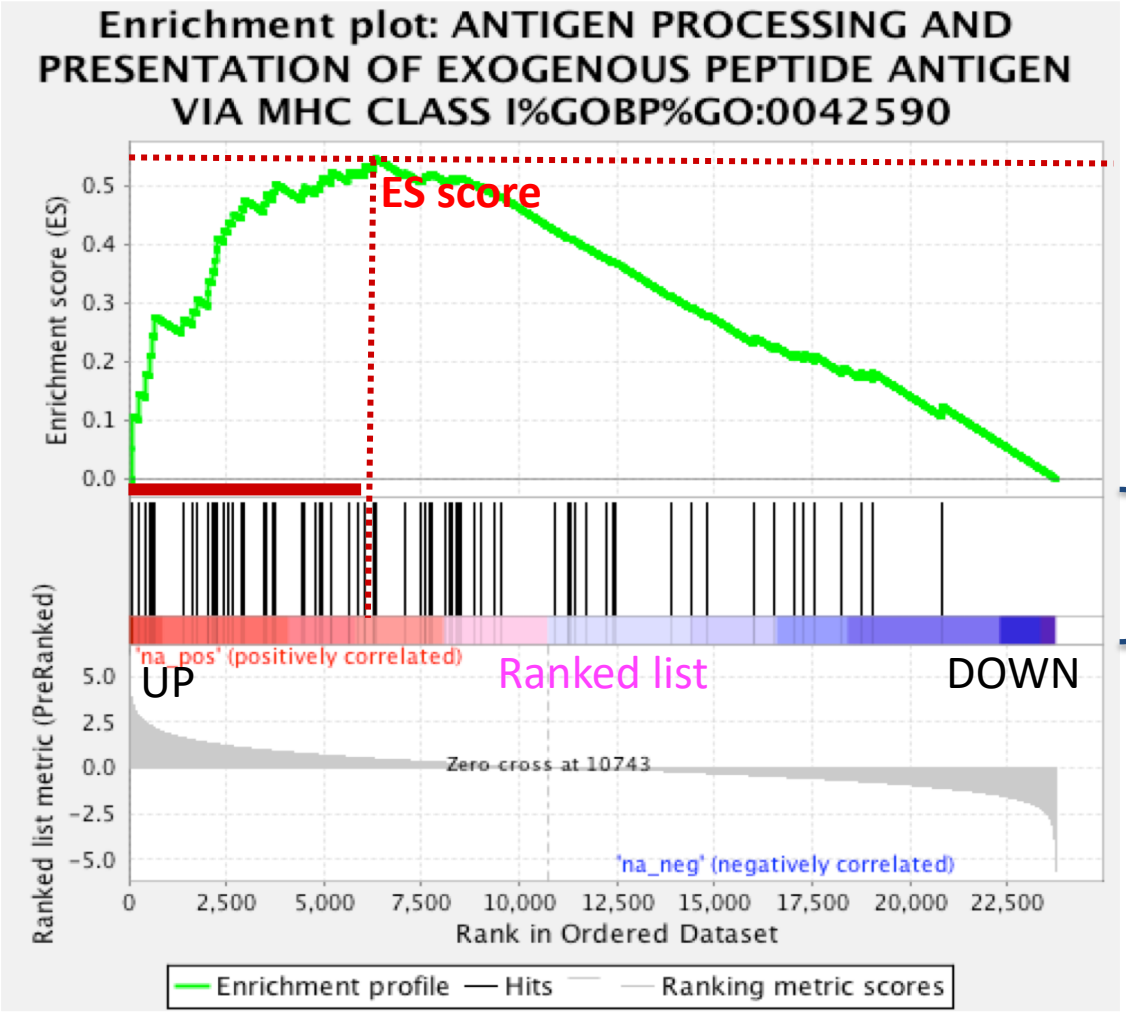
http://www.people.vcu.edu/~mreimers/HTDA/Mootha%20-%20GSEA.pdf

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

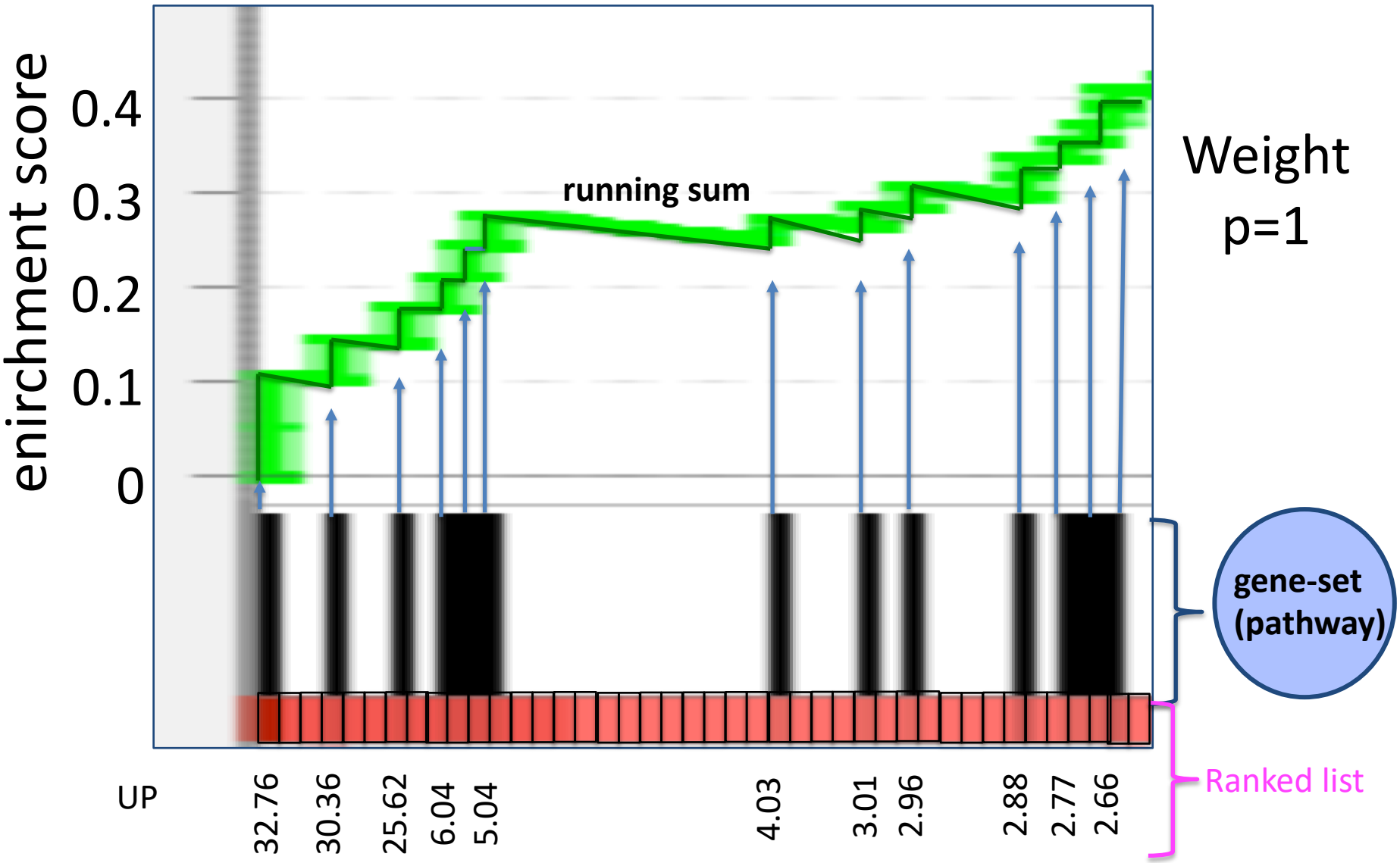
DOWN



gene-set (pathway)

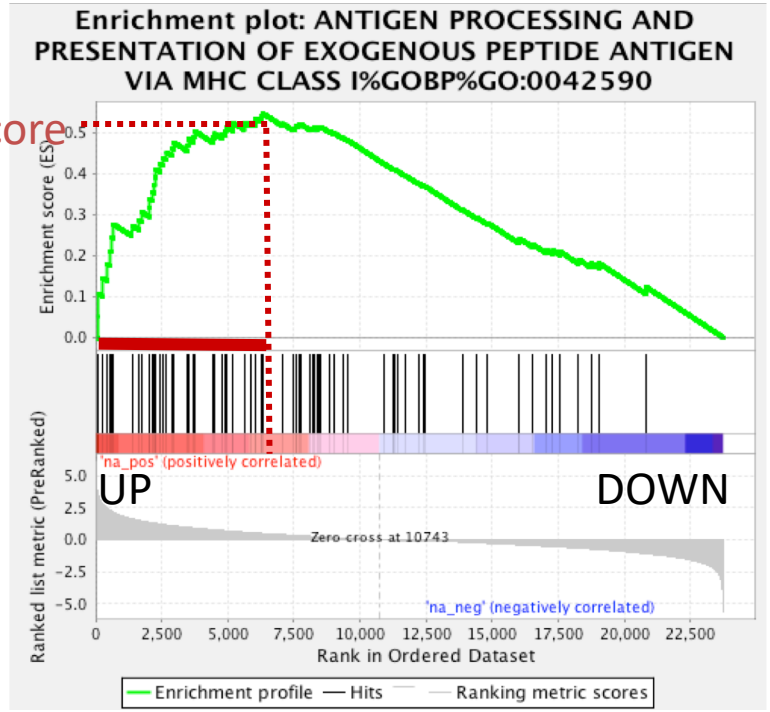
1. Maximum (or minimum) ES score is the final **ES score** for the gene set
2. Can define “leading edge subset” as all those genes ranked as least as high as the enriched set.

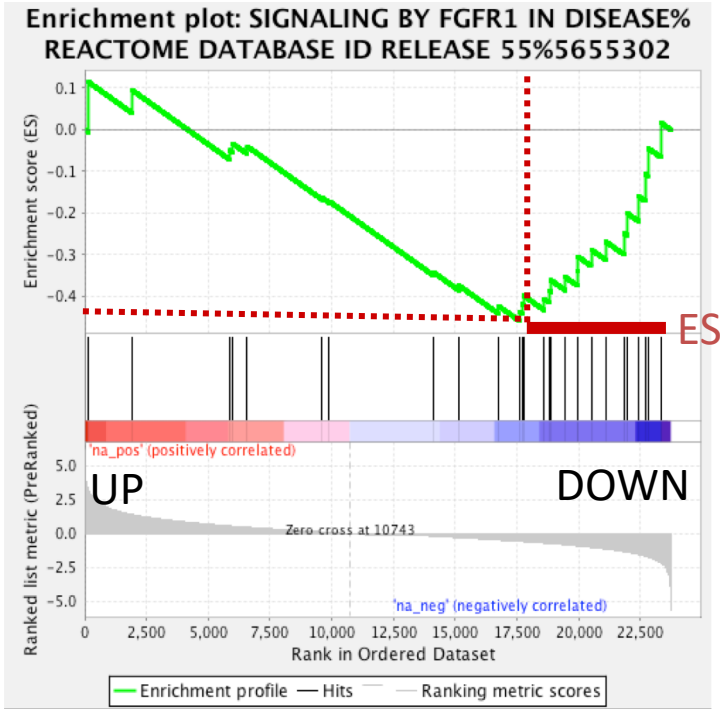
GSEA running sum



Positive and negative enrichment scores

ES score

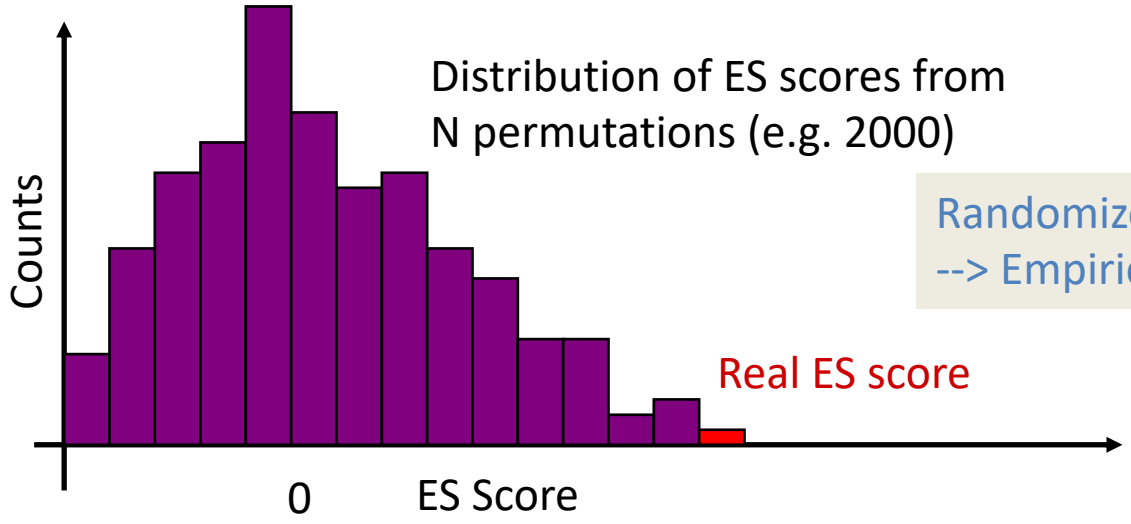




ES score

Going from ES score → P-value

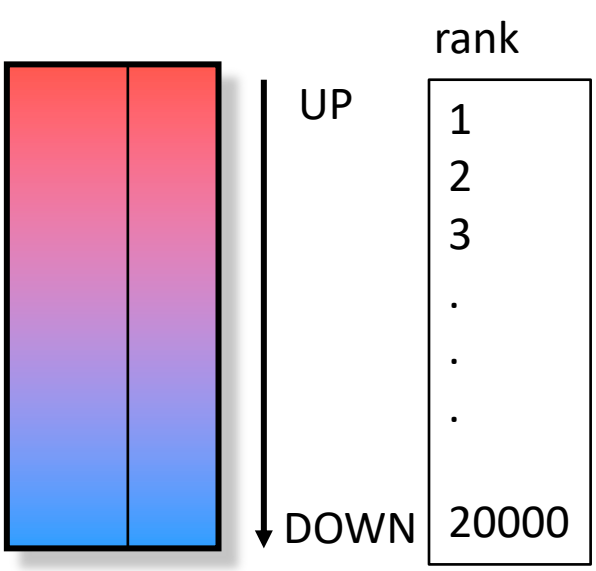
1. Generate null-hypothesis distribution from randomized data (see permutation settings)
2. Estimate empirical p-value by comparing observed ES score to null-hypothesis distribution from randomized data (for every gene-set)

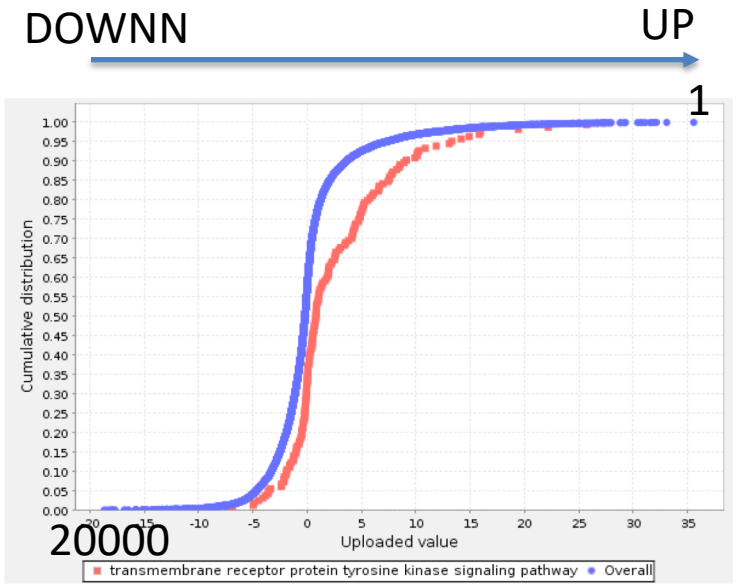


Randomized with ES score ≥ real: 4 / 2,000
--> Empirical p-value = 0.002

Other enrichment tests for a ranked gene list

Wilcoxon ranksum test





Panther

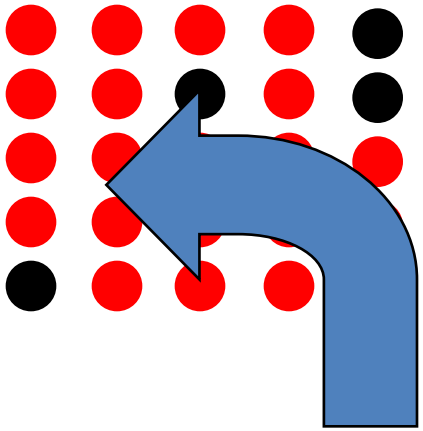
Outline of theory component

- 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

Multiple test corrections

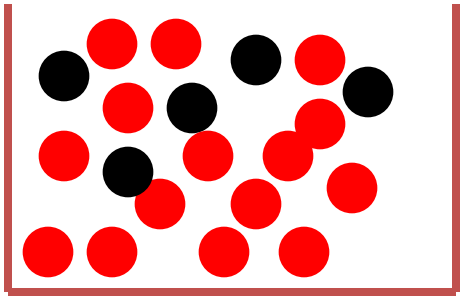
How to win the p-value lottery

Random draws

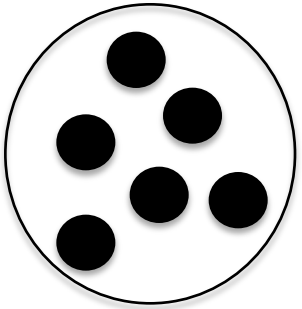


… 7,834 draws later …

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



Background population:
500 black genes,
4500 red genes



1 gene-set
(apoptosis)

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)*"**

Bonferroni correction caveats

- Bonferroni correction is very stringent and can "wash away" real enrichments leading to false negatives,
- Often one is willing to accept a less stringent condition, the "false discovery rate" (FDR), which leads to a gentler correction when there are real enrichments.

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"

False discovery rate (FDR)

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

2. **Adjusted P-value** is “nominal” P-value times # of tests divided by the rank of the P-value in sorted list: **P-value x [# of tests] / Rank**

3. **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**.

4. Look at which gene-sets have a **FDR of 0.05 or less** and report them as **significantly enriched.**

Benjamini-Hochberg example

Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional regulation*	0.001	0.001 x 53/1 = 0.053	0.040
2	*Transcription factor*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Initiation of transcription*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Nuclear localization*	0.0031	0.0031 x 53/4 = 0.040	0.040
5	*Chromatin modification*	0.005	0.005 x 53/5 = 0.053	0.053
…	…	…	…	…
52	*Cytoplasmic localization*	0.97	0.985 x 53/52 = 1.004	0.99
53	*Translation*	0.99	0.99 x 53/53 = 0.99	0.99

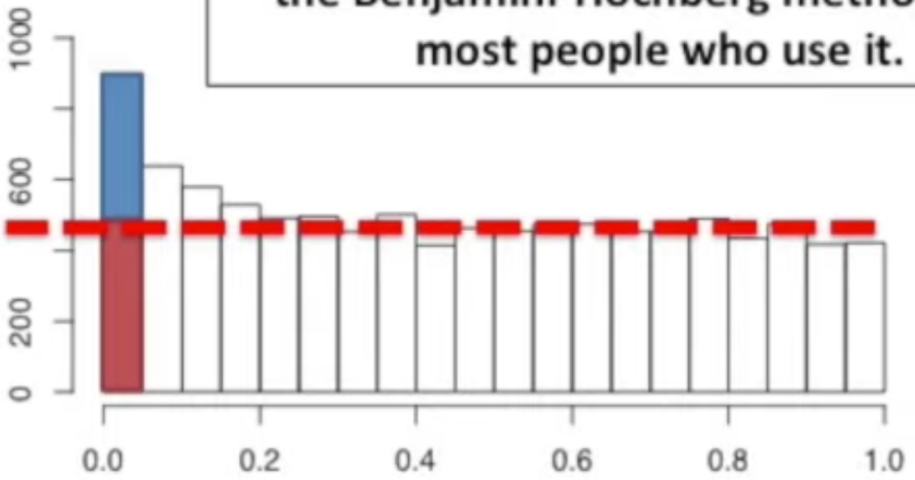
Q-value (or FDR) corresponding to a nominal P-value is the smallest adjusted P-value assigned to P-values with the same or larger ranks.

Gene set enrichment significant at FDR < 0.05

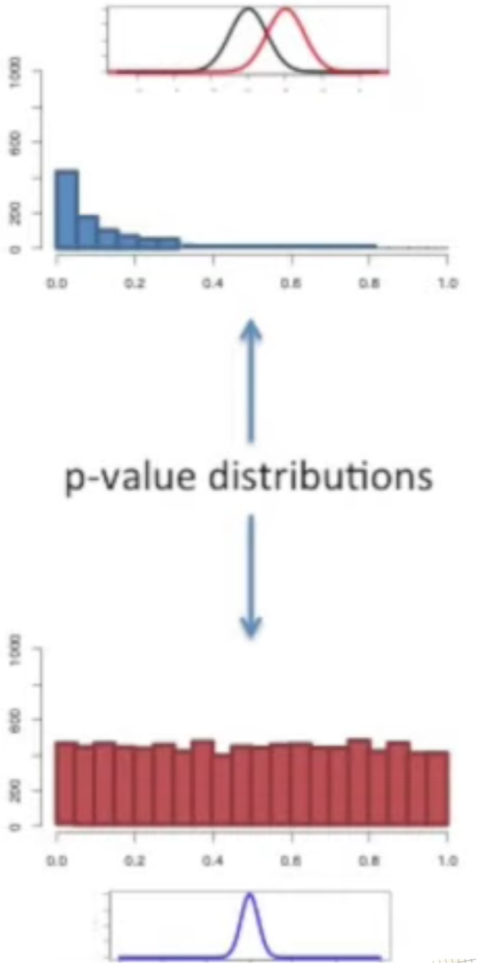
How to win the p-value lottery, part 2

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



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

Reducing multiple test correction stringency

- The correction to the P-value threshold α depends on the # of tests that you do, so, no matter what, the more tests you do, the more sensitive the test needs to be
- Can control the stringency by reducing the number of tests: e.g. use GO slim; restrict testing to the appropriate GO annotations; or filter gene sets by size.

Summary

Multiple test correction

- **Bonferroni**: stringent, controls probability of at least one false positive*
- **FDR**: more forgiving, controls expected proportion of false positives* -- typically uses Benjamini-Hochberg

* Type 1 error, aka probability that observed enrichment if no association

What Have We Learned?

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 hypothesi

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:[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 [illegible]%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 [illegible]%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 PR[illegible]GO:001	163	1.75	0	0.0550066
ATP-DEPENDENT RNA HELICASE ACTIVITY%G[illegible]GO:0004004	20	1.74	0.0059	0.0556722
ACID-AMINO ACID LIGASE ACTIVITY%G[illegible]O: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%[illegible]%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_1[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

TABULAR FORMAT

NETWORK VISUALIZATION

Many available enrichment analysis tools

web-based

Cytoscape app

Standalone

R package

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 ;

Ranked list

	GSEA	PANTHER
Rank test	Modified KS test	Wilcoxon Rank Sum test
Correction for multiple hypothesis testing	yes	yes
Possibility to visualize results with Cytoscape enrichment map	yes	no

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

We are on a Coffee Break & Networking Session

compute canada | calcul canada

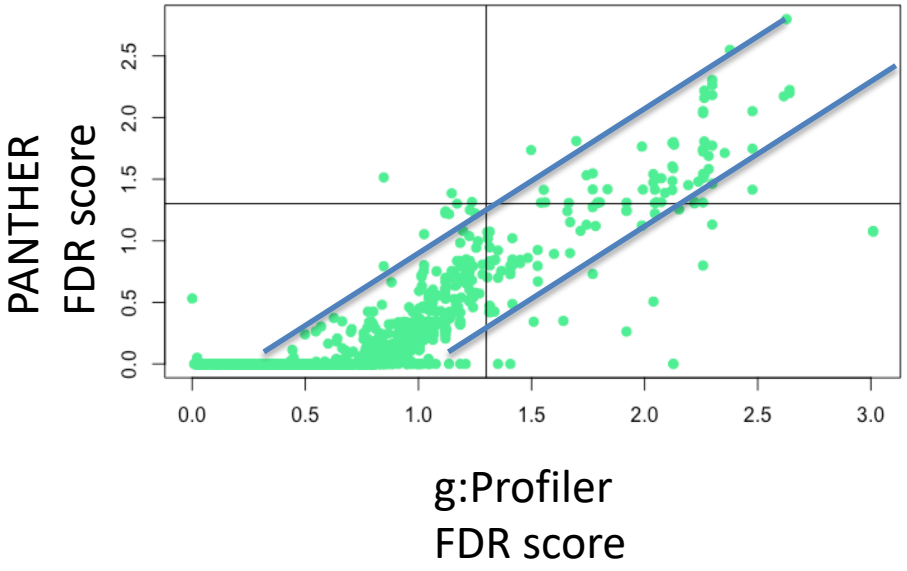
Workshop Sponsors:

Canadian Centre for Computational Genomics

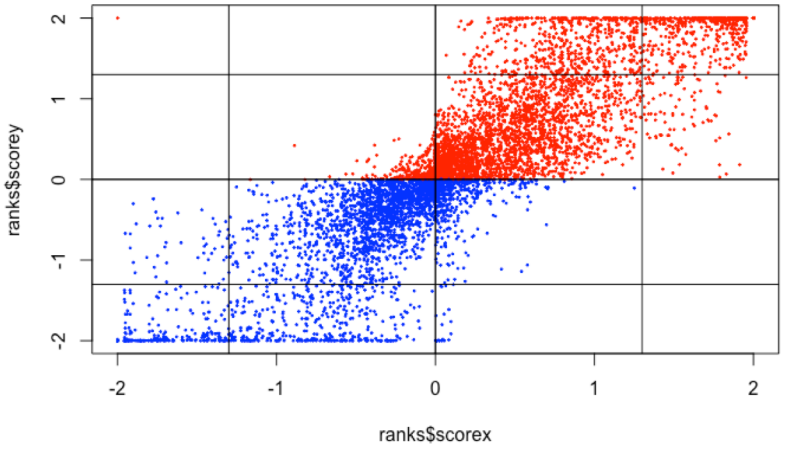
Additional slides

Comparison of results

gene list



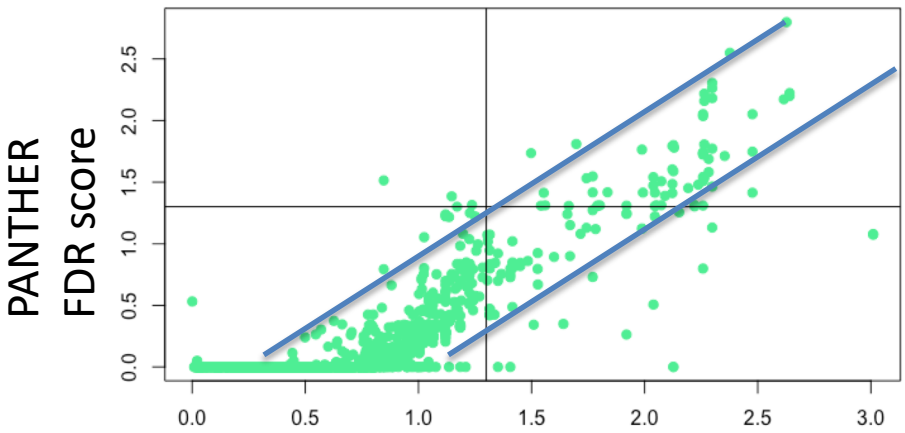
Ranked list

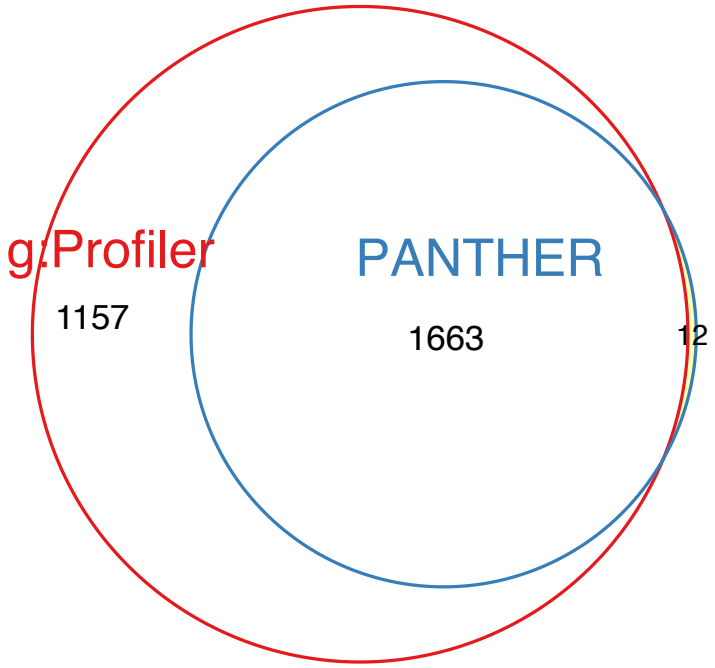


Similar results are obtained between g:Profiler and PANTHER

Comparison of results

gene list

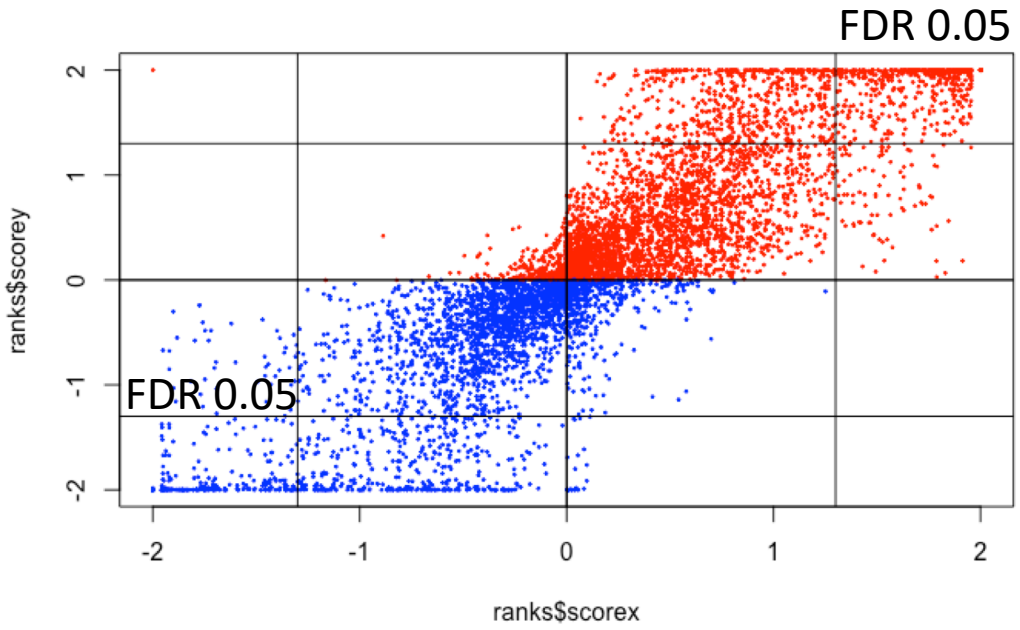


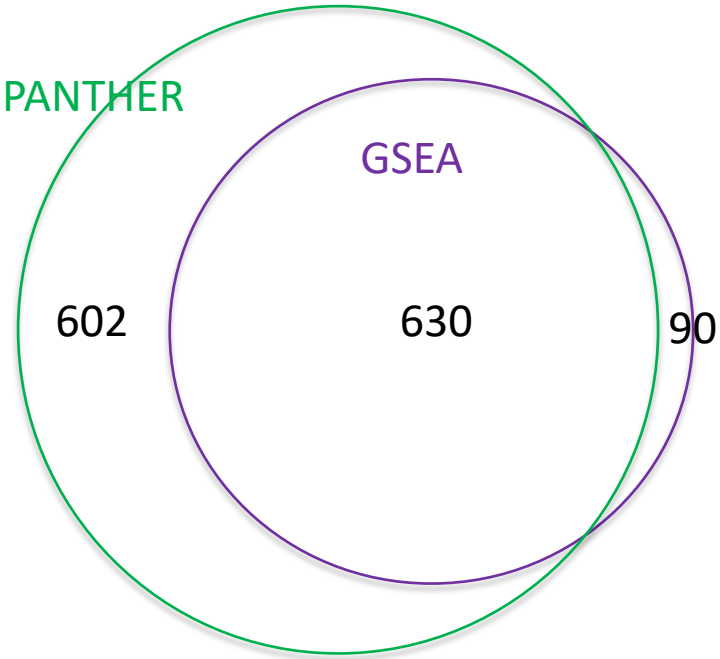


g:Profiler 16132 gene-sets
PANTHER 15815 gene-sets

Number of gene-sets significant under FDR <0.05

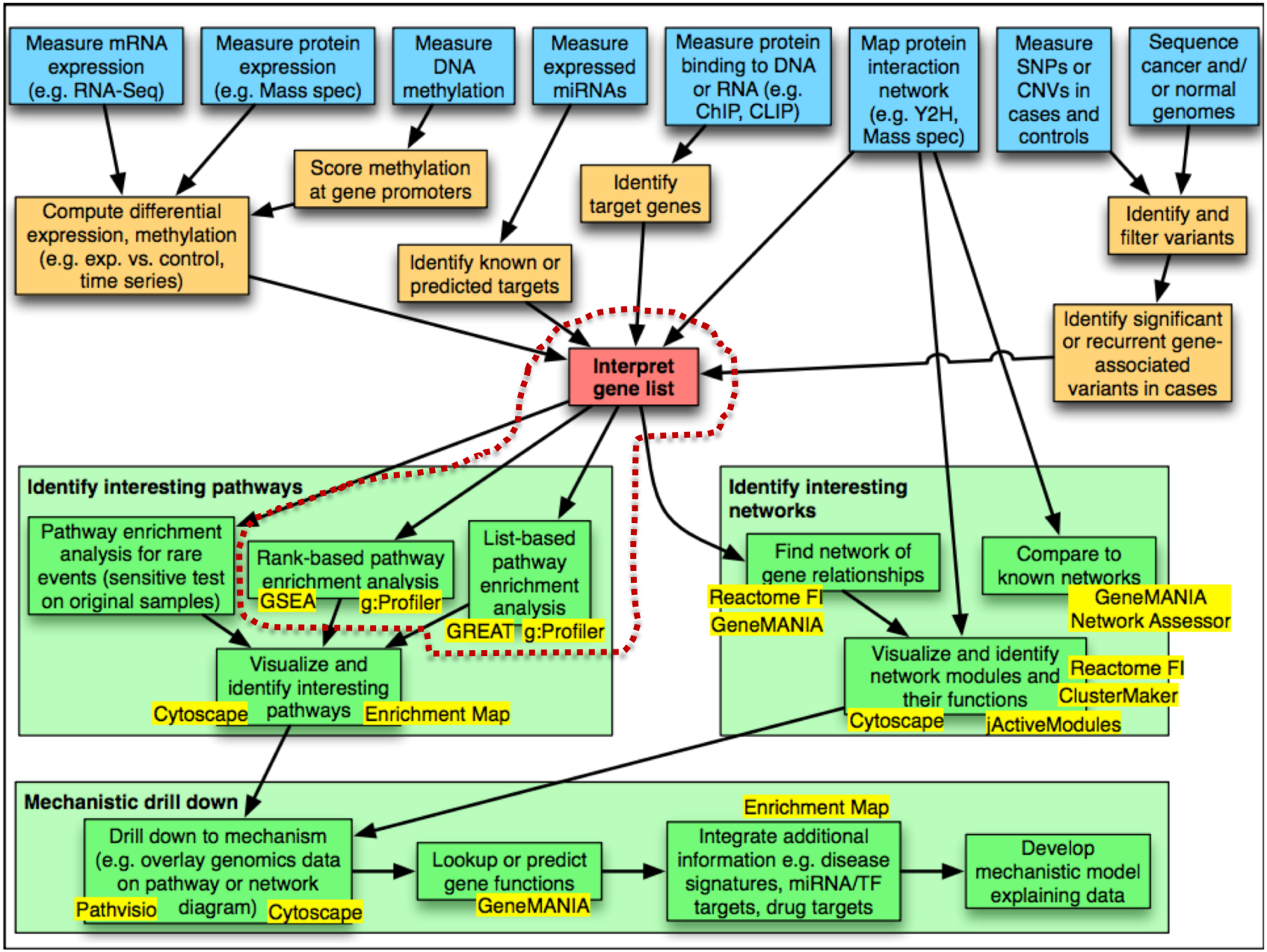
Ranked list



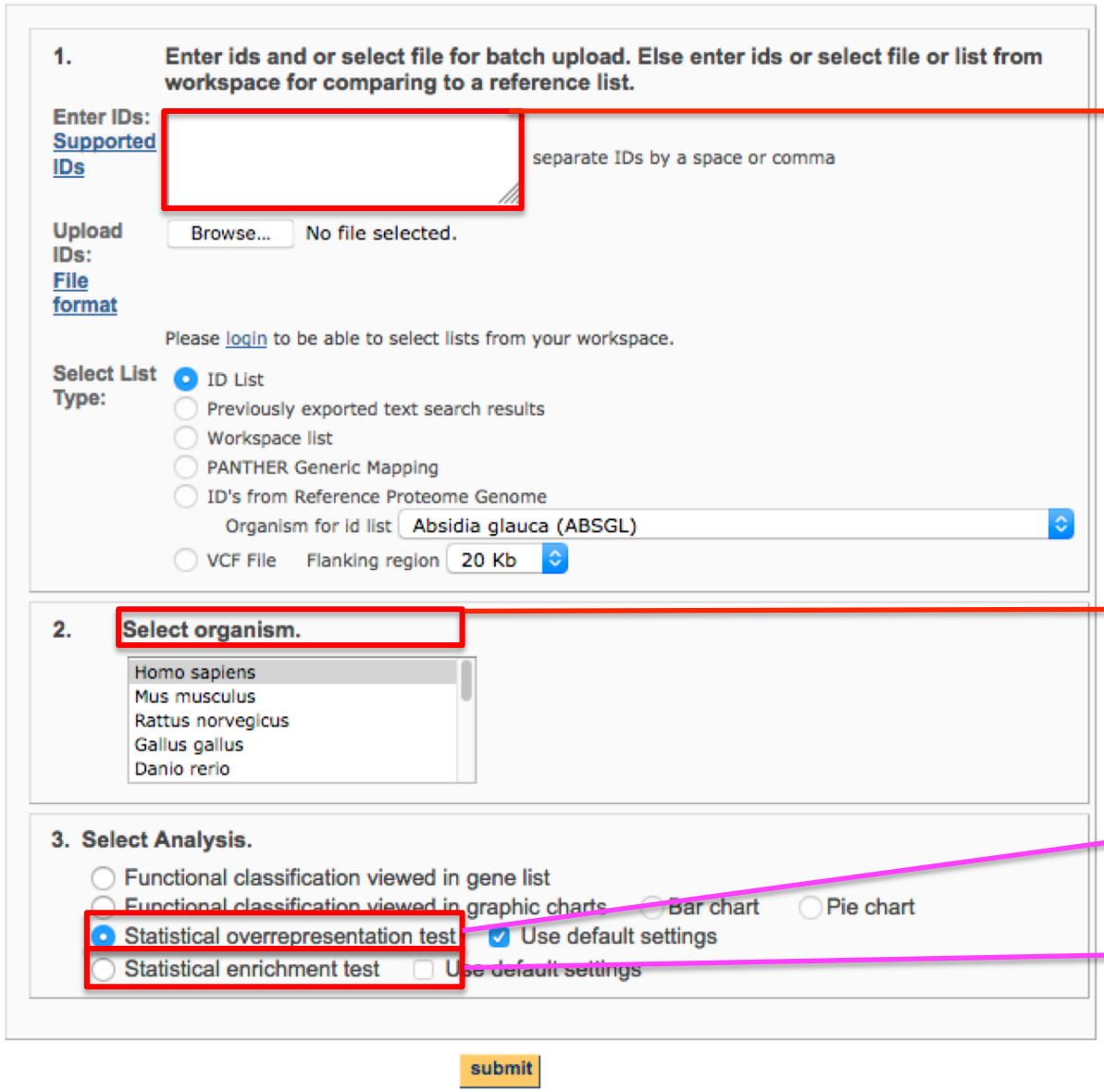


Number of gene-sets significant under FDR <0.05

- gene-set enriched in genes up-regulated
- gene-set enriched in genes down-regulated


Measure mRNA expression (e.g. RNA-Seq)
Measure protein expression (e.g. Mass spec)
Measure DNA methylation
Measure expressed miRNAs
Measure protein binding to DNA or RNA (e.g. ChIP, CLIP)
Map protein interaction network (e.g. Y2H, Mass spec)
Measure SNPs or CNVs in cases and controls
Sequence cancer and/ or normal genomes
Score methylation at gene promoters
Identify target genes
Identify and filter variants
Compute differential expression, methylation (e.g. exp. vs. control, time series)
Identify known or predicted targets
Identify significant or recurrent gene-associated variants in cases
Interpret gene list
Identify interesting pathways
Pathway enrichment analysis for rare events (sensitive test on original samples)
Rank-based pathway enrichment analysis
GSEA
g:Profiler
List-based pathway enrichment analysis
GREAT
g:Profiler
Visualize and identify interesting pathways
Cytoscape
Enrichment Map
Identify interesting networks
Find network of gene relationships
Reactome FI
GeneMANIA
Compare to known networks
GeneMANIA
Network Assessor
Visualize and identify network modules and their functions
Cytoscape
Reactome FI
ClusterMaker
jActiveModules
Mechanistic drill down
Drill down to mechanism (e.g. overlay genomics data on pathway or network diagram)
Pathvisio
Cytoscape
Lookup or predict gene functions
GeneMANIA
Enrichment Map
Integrate additional information e.g. disease signatures, miRNA/TF targets, drug targets
Develop mechanistic model explaining data

PANTHER (website: http://pantherdb.org/)



Input gene list format:
Gene list
Vcf file →Gene list
Ranked gene list

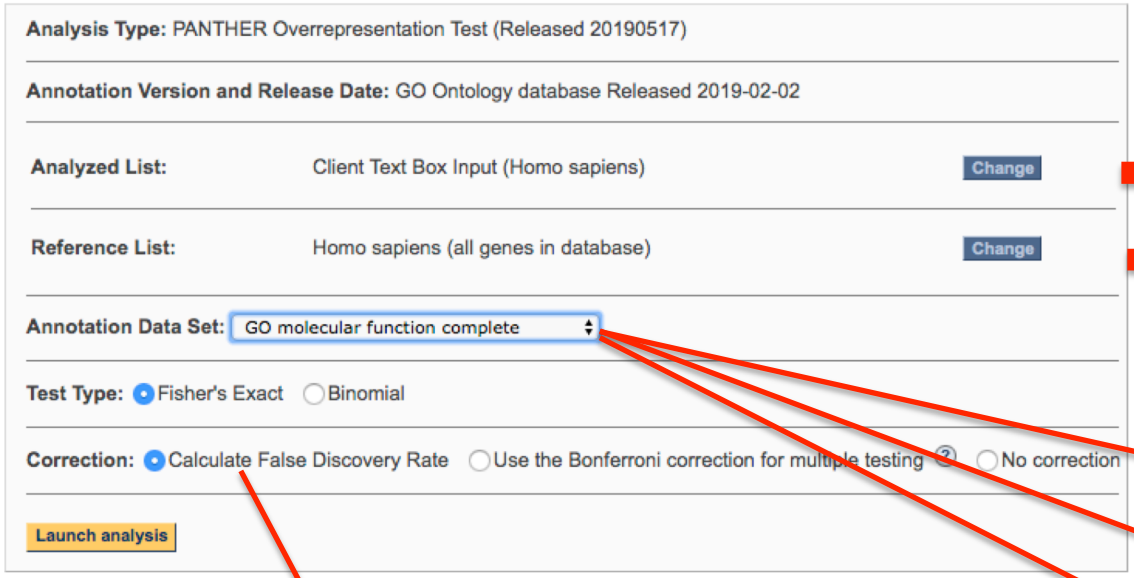
131 model organisms

Fisher ‘s exact test (thresholded gene list)

Wilcoxon rank test (rank list)

PANTHER (website)

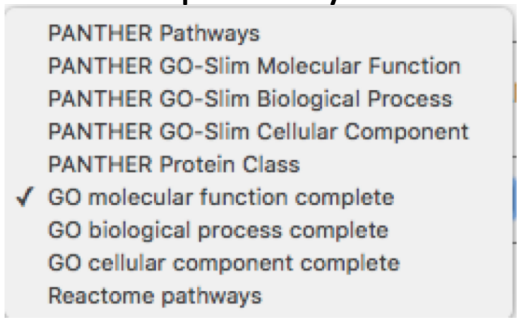
Over-representation Analysis (Fisher's exact test)



thresholded gene list

background/universe

Choose pathway database



Correction for multiple hypothesis testing

Updated frequently!!

PANTHER output

of genes in original pathway

# of genes in my gene list and tested pathway

Significance of the enrichment after correction for multiple hypothesis testing.

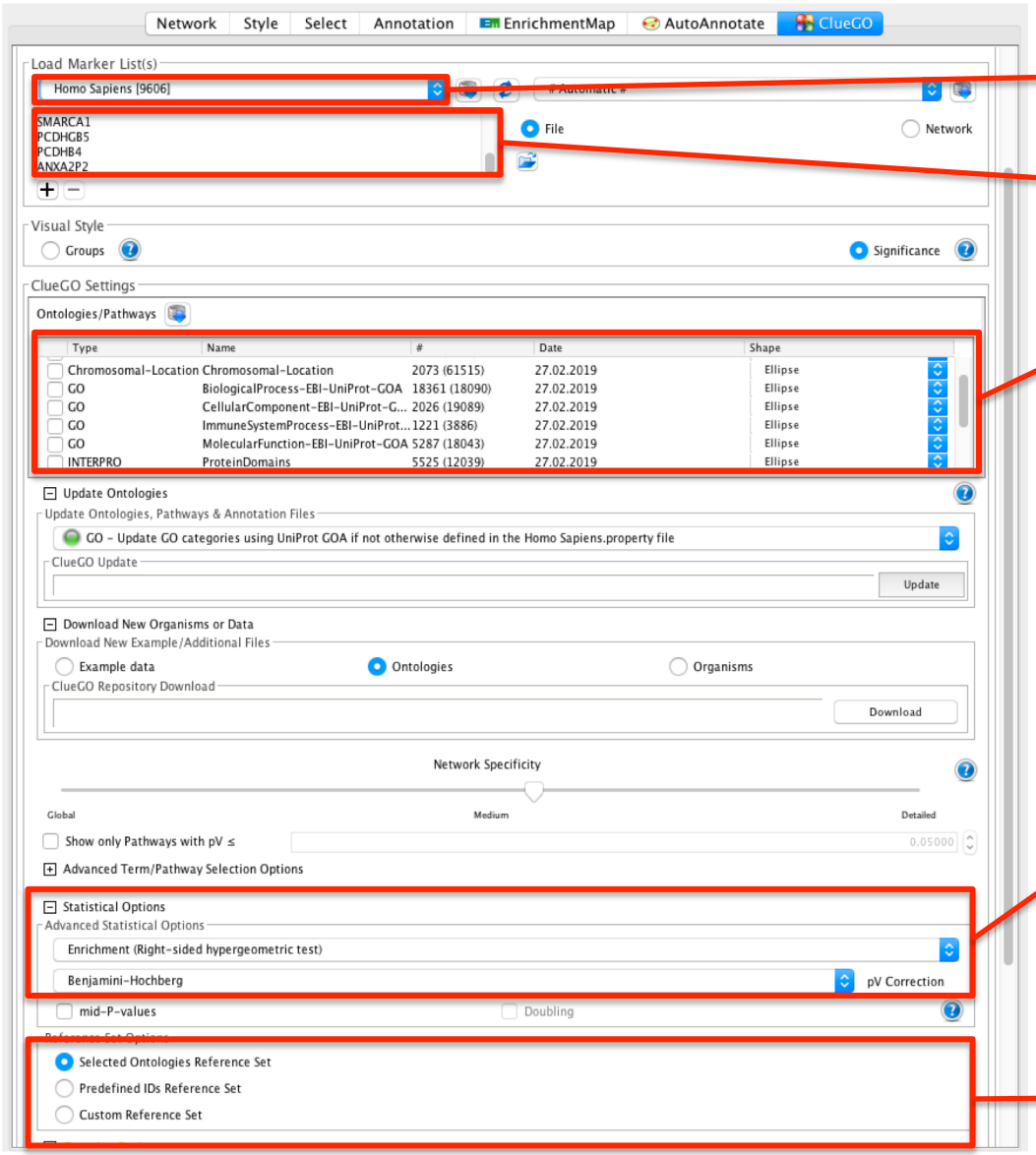
Pathway (gene-sets)

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	Δ +/-	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

Only text output visualization?

ClueGO (Cytoscape app)



Selection of model organism (Homo sapiens or Mus musculus)

Input format: thresholded gene list

Pathway database selection (my selection): **(regularly updated)**

* GO BP
* Reactome Pathway

Statistical test: Fisher's exact test

my selection:

Overrepresentation analysis

FDR for multiple hypothesis correction

Upload a background if needed

EnrichR

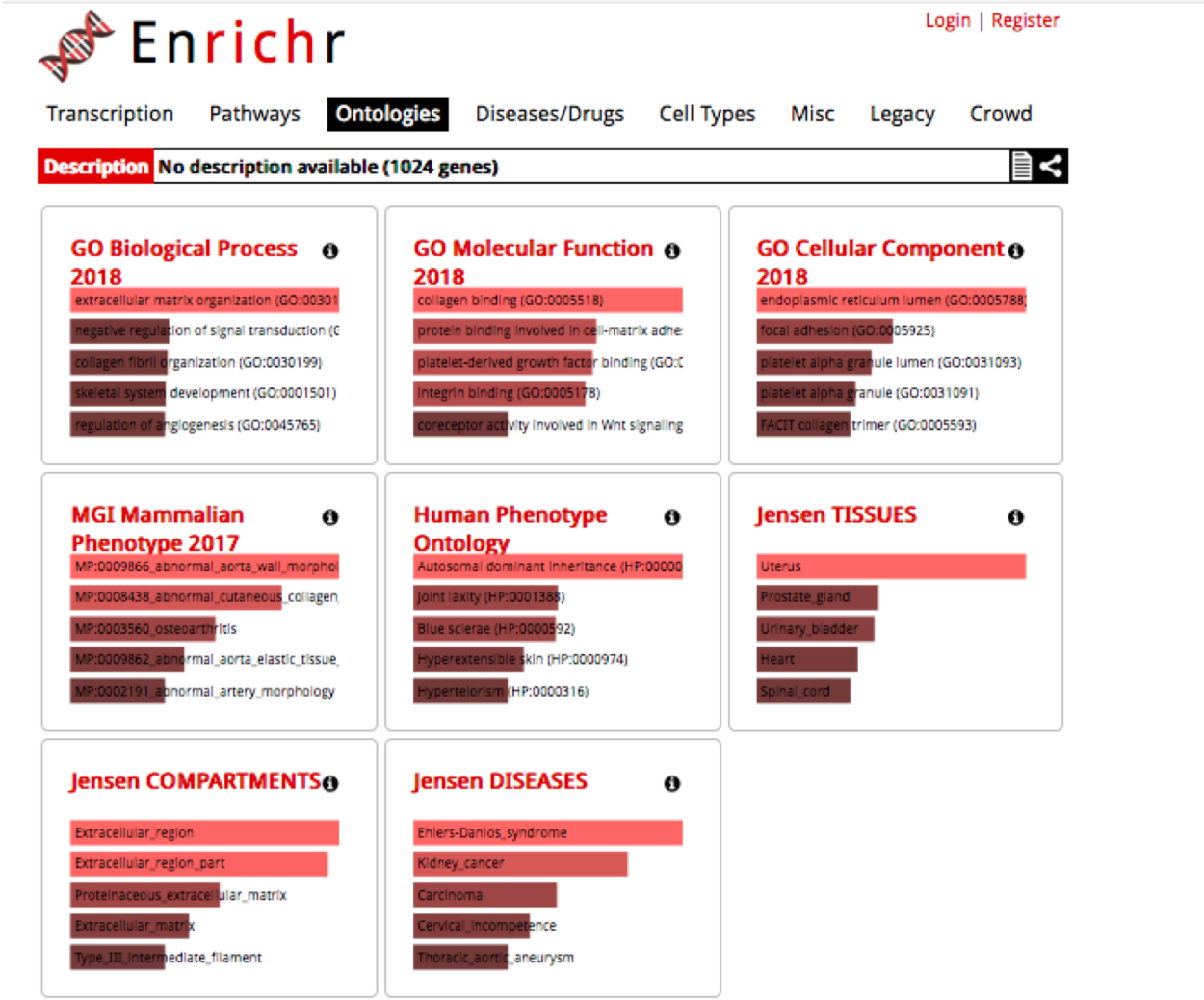
Input

Gene list
or Bed file

Test:
Fisher's exact test,
pvalue corrected
for multiple hypothesis
testing

Output:
Table or grahs

? No option to put
Reference background (use
only if you are doing a
whole genome study)



Enrichr output table

GO Biological Process 2018

Need to select positive score for overrepresentation

Fisher's exact test

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;TV
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)

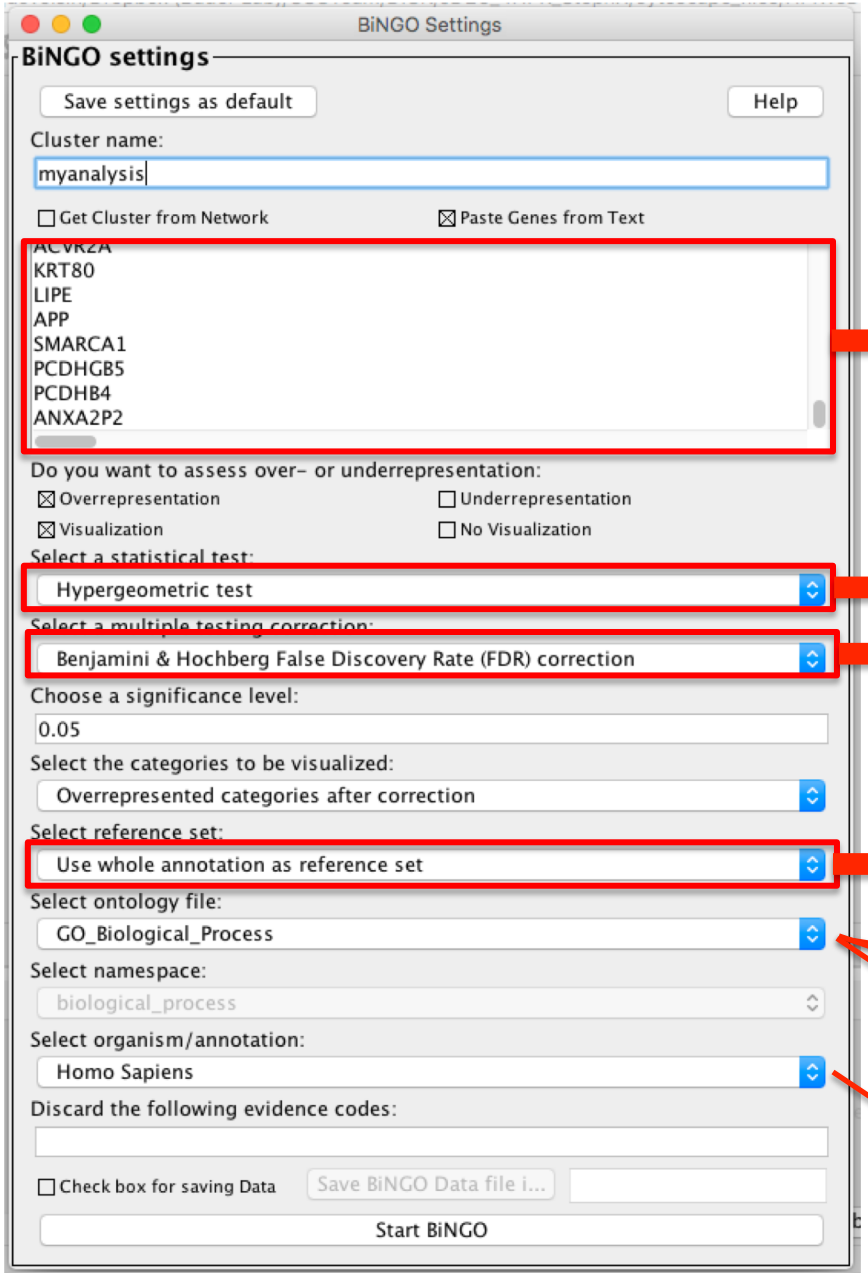
Overlap:
Numerator -> genes in my gene list and tested pathway

Denominator -> Genes in the original pathwyay

FDR:
Correction for multiple hypothesis testing

List of genes in the overlap

BiNGO (Cytoscape app)



thresholded gene list

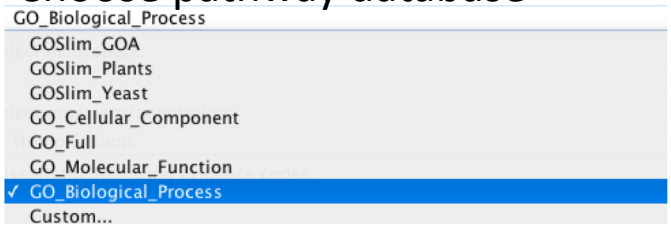
Fisher's exact test

Correction for multiple hypothesis testing

Are the databases updated frequently? If not, we should **use custom gmt option**

background/universe

Choose pathway database



Different model organisms

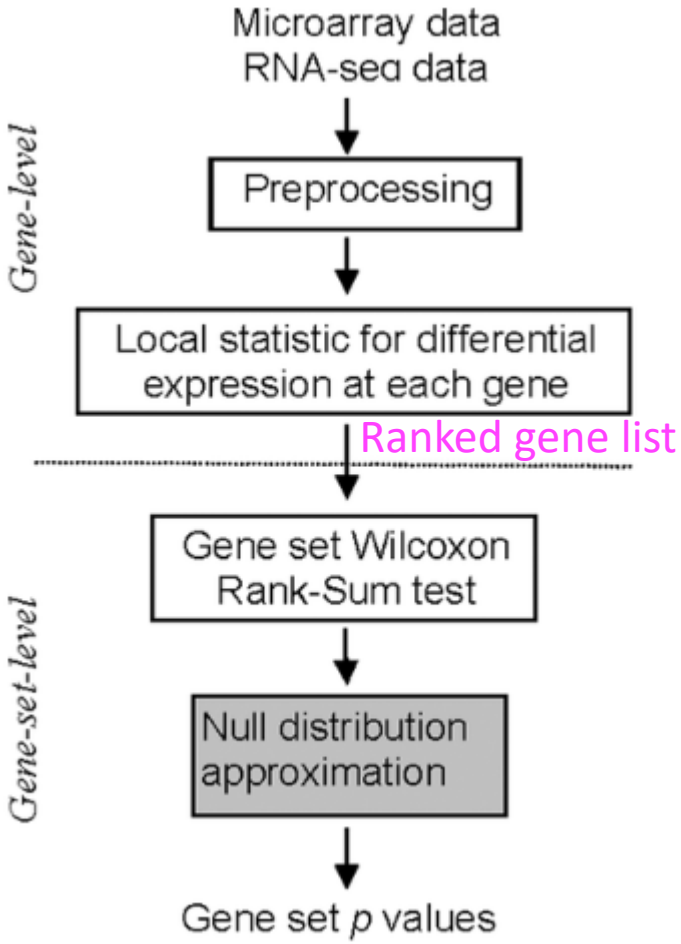
BiNGO output

1. **2.** **3.** **4.** **5.**

GO-ID	Description	p-val	corr p-val	cluster freq	total freq	genes
48731	system development	8.7490E-55	3.1041E-51	316/805 39.2%	2416/14265 16.9%	SPON2 ERRFI1 APP SPARC SERPINE1 COL12A1 XYLT1 STMN3 ELK3 AQP1 NDST1 ...
48856	anatomical structure development	1.1175E-52	1.9824E-49	330/805 40.9%	2649/14265 18.5%	SPON2 ERRFI1 APP SPARC SERPINE1 COL12A1 XYLT1 STMN3 ANTXR1 ELK3 AQP1...
7275	multicellular organismal development	6.5566E-52	7.7543E-49	352/805 43.7%	2965/14265 20.7%	SPON2 ERRFI1 APP SPARC SERPINE1 COL12A1 DIXDC1 XYLT1 STMN3 ELK3 AQP1 ...
32502	abelFor=,text=developmental process,verticalAlignment=CENTER,verticalTextPosition=CENTER]	5.2369E-49	4.6452E-46	365/805 45.3%	3227/14265 22.6%	SPON2 ERRFI1 APP SPARC SERPINE1 COL12A1 DIXDC1 XYLT1 STMN3 ANTXR1 EL...
7155	cell adhesion	1.9208E-47	1.3630E-44	149/805 18.5%	711/14265 4.9%	SEMA5A SPON2 APP COL16A1 COL12A1 ANTXR1 CTGF LOXL2 COMP CDH5 ISLR ...
22610	biological adhesion	2.3126E-47	1.3675E-44	149/805 18.5%	712/14265 4.9%	SEMA5A SPON2 APP COL16A1 COL12A1 ANTXR1 CTGF LOXL2 COMP CDH5 ISLR ...
9653	anatomical structure morphogenesis	7.9139E-38	4.0112E-35	184/805 22.8%	1214/14265 8.5%	SEMA5A SPON2 ERRFI1 APP SERPINE1 FGF1 ANTXR1 CTGF ELK3 AQP1 COMP NDS...
48513	organ development	9.3874E-37	4.1633E-34	231/805 28.6%	1788/14265 12.5%	ERRFI1 APP SERPINE1 ELK3 AQP1 NDST1 GJA1 EDNRA KDR HOXA3 HOXA1 SOX7 S...
1944	vasculature development	6.7888E-32	2.6763E-29	75/805 9.3%	273/14265 1.9%	SEMA5A ERRFI1 ROBO4 TCF21 SHB FGF1 GLI3 CYR61 CTGF ELK3 CDH5 GJA1 EDN...
1568	blood vessel development	2.6225E-30	9.3046E-28	72/805 8.9%	265/14265 1.8%	SEMA5A ROBO4 SHB FGF1 GLI3 CYR61 CTGF ELK3 CDH5 GJA1 EDNRA PLAU KDR ...
32501	multicellular organismal process	6.4334E-30	2.0750E-27	396/805 49.1%	4368/14265 30.6%	SERPINE1 COL12A1 XYLT1 STMN3 ELK3 NDST1 GPR176 GJA1 DPYSL4 DPYSL3 HO...

- **1**. each row is a pathway (gene-set) that was in the original pathway database that we selected (GO Biological Process)
- **2**. corr pval: the most important column as it is the pvalue corrected for multiple hypothesis.
- **3** .information about the size of the overlap between my gene list and and the pathway (gene-set).
- **4**. information about the size of the original pathway (gene-set) in the chosen pathway database (GO biological process).
- **5**. Genes in the my gene list and in the pathway (gene-set).

Wilcoxon rank sum test (Mann-Whitney U Test)



- Rank based **non parametric test** for comparing two groups of observations without the assumption of certain distributions.
- It has been implemented in many packages and software for gene-set testing (limma R function geneSetTest, R package SAFE using safe(), Gostat(), Panther.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031505

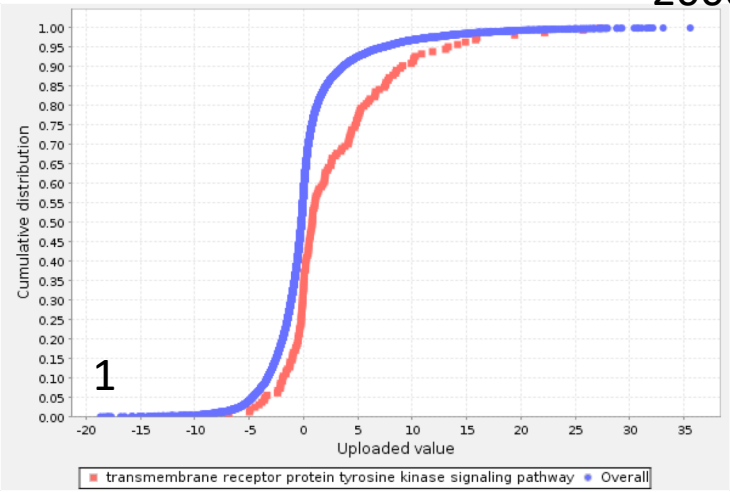
Wilcoxon rank sum test

(as used and described in PANTHER tool -pantherdb.org/)

Step 1 out of 3

- **All genes rank sum**:
  - All genes ordered by expression values to create a rank.
  - Genes with the smallest values get a rank of 1.
  - A rank sum is calculated (summing up the ranks for all genes)
  - The average rank, R2, is calculated by dividing the rank sum by the total number of genes uploaded, n2.

20000



Rank sum = 1 + 2 +3 +….+ 20000

n2 = 20000

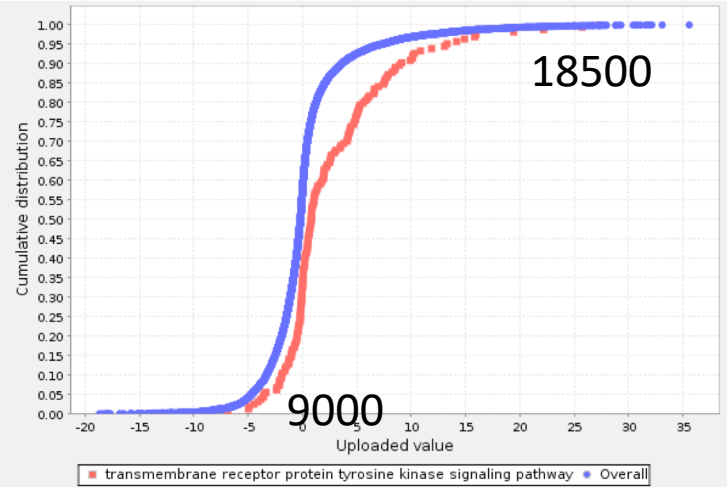
R2 = rank sum / n2

Wilcoxon rank sum test

(as used and described in PANTHER tool -pantherdb.org/)

Step 2 out of 3

- **Gene set rank sum**:
  - A rank sum calculated for genes in the tested gene-set: sum up the ranks for all genes that map the gene-set.
  - The average rank, R1 is then calculated by dividing the rank sum by the number of genes, n1, that map to the category.



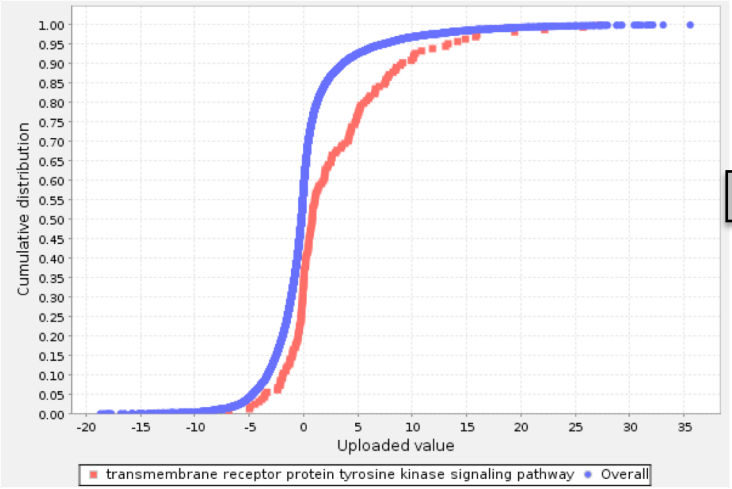
Rank sum = 9000 + 9005 + …18500
n1 = 250 (size of gene set)
R1 = rank sum / n1

Wilcoxon rank sum test

(as used and described in PANTHER tool -pantherdb.org/)

Step 3 out of 3

- **The Mann Whitney U statistic is calculated :**
  - U1 = n1* n2 + (n1 * (n1 + 1)) / 2 - R1 (gene-set)
  - U2 = n2* n2 + (n1 * (n2 + 1)) / 2 - R2 (all genes)
  - **U**: The larger of these two values is the **Mann Whitney U-statistic**,
  - **Pvalue** associated with U, whose distribution for small sample sizes can be found in most statistic books or use the normal approximation (Z-score = (U- (n1* n2)/2)/sqrt(n1*n2*(n1+n2+1)/12).



- The distribution of values for your uploaded list is shifted towards greater values than the overall distribution of all genes that were uploaded.
- A small, significant p-value indicates that the distribution for this category is non-random and different than the overall distribution. A cutoff of 0.05 is recommended as a starting point.

Minimum hypergeometric test (mHG)

(used in g:Profiler, ordered query)

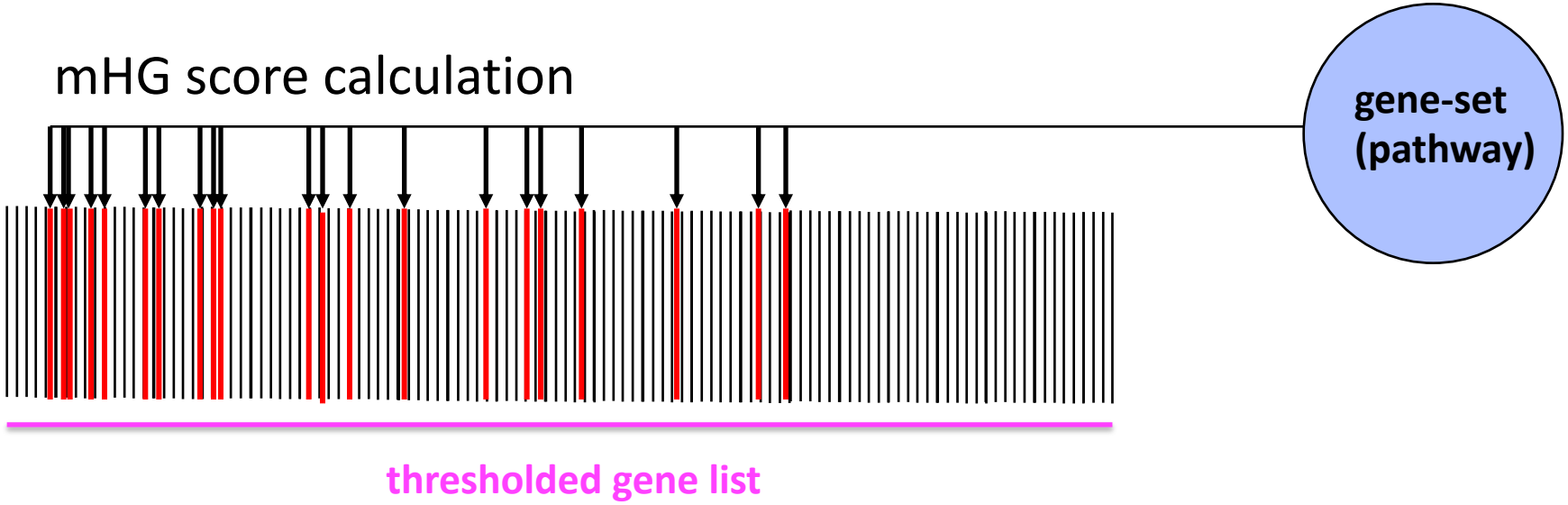
Steps

1. Calculate p-value at multiple thresholds

1. Correct for multiple testing (or compute empirical p-values using permutations)

Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 2007 Mar 23;3(3):e39

mHG Method



Where are the gene-set genes located in the ranked list?
Is there distribution random, or is there an enrichment in either end?

Gene list (threshold) chosen is the one associated with the lowest pvalue