廣州醫科大學
GUANGZHOU MEDICAL UNIVERSITY

Gene Set Analysis –Methods and Tools

Antonio Mora, Ph.D.

(Wechat: antoniocmora)

20.12.2018

MORA LAB
FUNCTIONAL BIOINFORMATICS

Contents

2.1. What is Gene Set Analysis?

You may have heard about:

- Pathway (enrichment) Analysis
- Gene Set (enrichment) Analysis
- Functional Enrichment Analysis
- Ontology Analysis
- Knowledge-driven pathway analysis
- And other names…

It is all the same. We are at the end of a research project and we want to find the meaning of the group of biological molecules that we obtained as a result. What is interesting about them? How are they related to each other?

Question: What is interesting about a group of genes?

Simplest method:
Google/Baidu/Pubmed your gene and read the papers.

Gene set analysis: Interpreting the query set as pathways or other gene sets.

- Results easier to interpret (familiar processes),
- Mechanistic (suggests possible mechanisms),
- Statistics taking into account.

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

"Gene Set Analysis" Elements:

A query set: A group of genes that were the result of some experiment

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

Reference Databases: Pathway / Ontology / Gene set DBs.

KEGG Kyoto Encyclopedia of Genes and Genomes

REACTOME A CURATED PATHWAY DATABASE

BETA WIKIPATHWAYS Pathways for the People

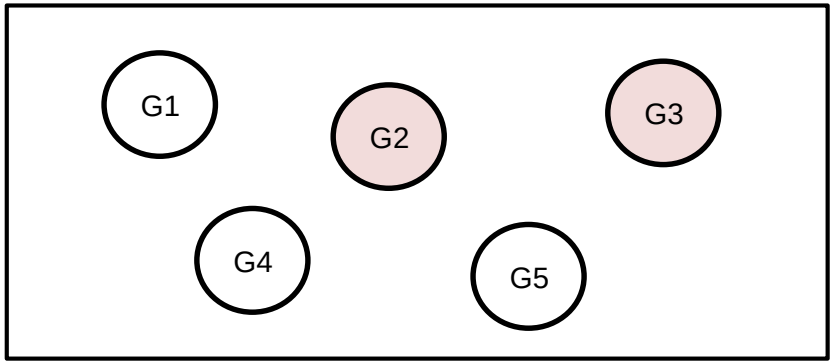
Statistical Method

Is my group of genes more enriched in one specific gene set than a group of genes randomly chosen?

"Gene Set Analysis" Questions:

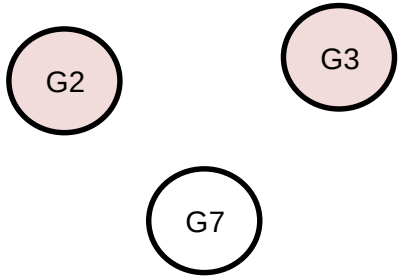
1. The ORA / Gene List approach: (e.g. genes with expression change > 2-fold)

Pathway X:



Answers the question:
Is Pathway X "surprisingly enriched" with genes from my gene list?

My Gene List:

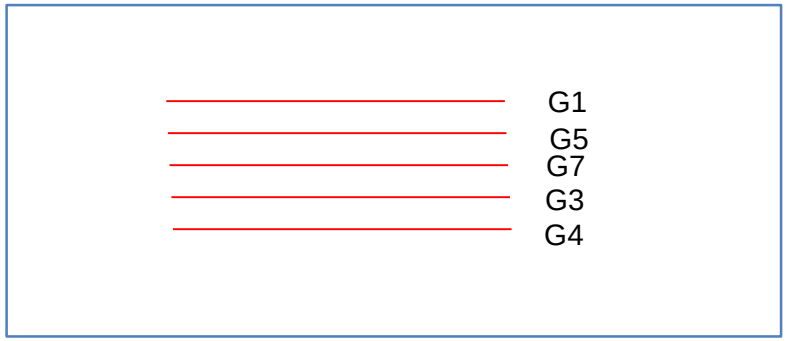


Or, in general (for the entire database):
Are any gene sets (pathways, complexes, diseases, functions) surprisingly enriched with genes from my gene/transcript list?

"Gene Set Analysis" Questions:

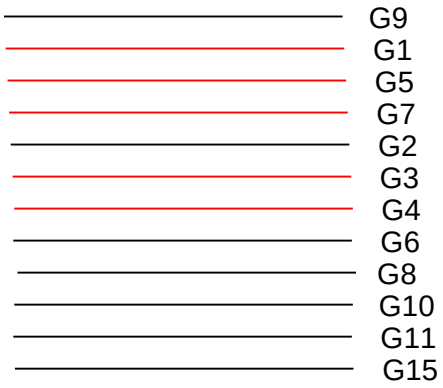
2. The FCS / Gene Rank approach: (e.g. entire list, ordered by differential expression)

Pathway X:



Answers the question:
Is Pathway X ranked "surprisingly high" when located on my ranked gene/transcript list?

My Gene Rank:



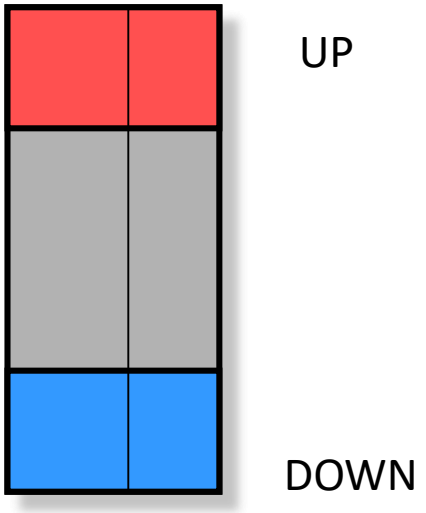
Or, in general (for the entire database):
Are any gene sets (pathways, complexes, diseases, functions) ranked surprisingly high when located on my ranked gene/transcript list?

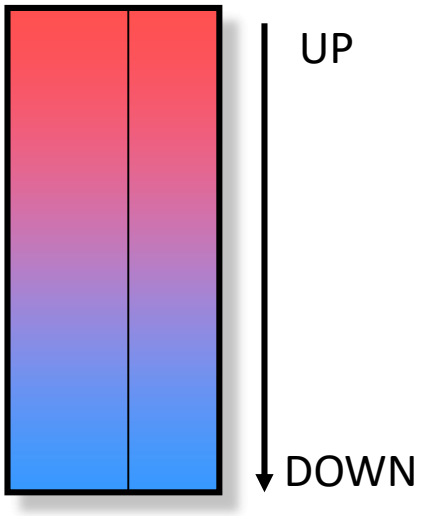
Statistical Tests

ENRICHMENT TEST

ORA / Gene list
Fisher's Exact Test (Hypergeometric), Binomial and Chi-squared.

FCS / Ranked list
GSEA,
Wilcoxon ranksum,
Mann-Whitney U,
Kolmogorov-Smirnov





Taken from: Canadian Bioinformatics Workshop

2.2. Before starting Gene Set Analysis

Where do query sets come from?:
- Molecular profiling (mRNA, protein)
- Interactions (TF binding sites, miRNA targets)
- Association studies (SNPs, CNVs)

Gene Set Enrichment Analysis:

Input: Query Set and Pathway/GeneSet database.

Example of query set: Differentially expressed genes (up-reg, down-reg, or the entire list).

The Gene / Protein List

- Be careful about gene/protein identifiers.
- Identifiers (IDs) are ideally unique, stable names or numbers that help track database records. For example, your wechat ID, Entrez Gene ID 41232, etc
- Gene and protein information stored in many databases
  - → Genes have many IDs
- Records for: Gene, DNA, RNA, Protein
  - Important to recognize the correct record type

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

We need both the query set and the pathways/gene sets using the same type of identifiers

Common Identifiers

Gene
Ensembl ENSG00000139618
Entrez Gene 675
Unigene Hs.34012

RNA transcript
GenBank BC026160.1
RefSeq NM_000059
Ensembl ENST00000380152

Protein
Ensembl ENSP00000369497
RefSeq NP_000050.2
UniProt BRCA2_HUMAN or A1YBP1_HUMAN
IPI IPI00412408.1
EMBL AF309413
PDB 1MIU

Species-specific
HUGO HGNC BRCA2
MGI MGI:109337
RGD 2219
ZFIN ZDB-GENE-060510-3
FlyBase CG9097
WormBase WBGene00002299 or ZK1067.1
SGD S000002187 or YDL029W

Annotations
InterPro IPR015252
OMIM 600185
Pfam PF09104
Gene Ontology GO:0000724
SNPs rs28897757

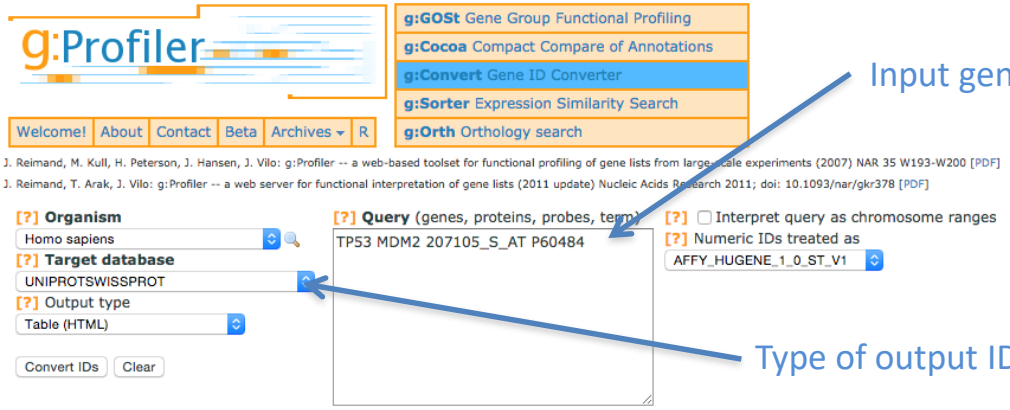
Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

Red = Recommended

Identifier Mapping

- So many IDs!
  - Software tools recognize only a handful
  - May need to **map** from your gene list IDs to standard IDs
- Four main uses
  - Searching for a favorite gene name
  - Link to related resources
  - Identifier translation
    - E.g. Proteins to genes, Affy ID to Entrez Gene
  - Merging data from different sources
    - Find equivalent records

ID Mapping Services

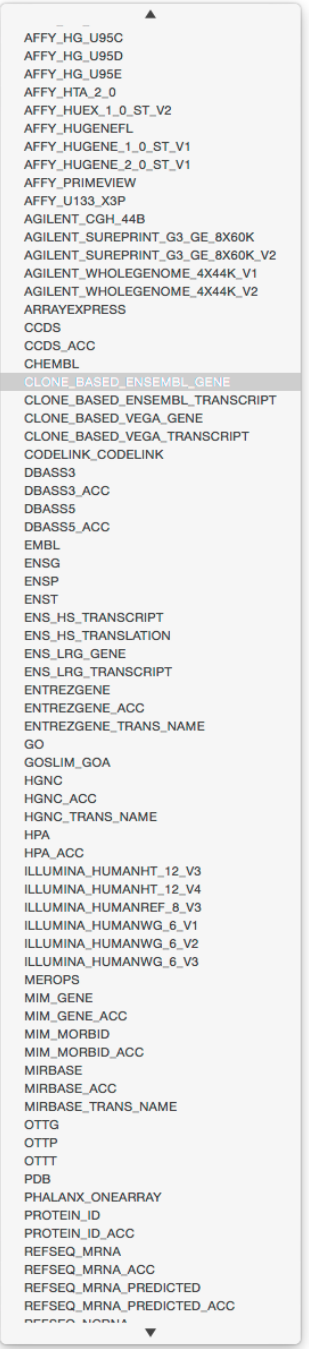


Input gene/protein/transcript IDs (mixed)

Type of output ID

>> Static URL
Come back later

g#	initial alias >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa	c#	converted alias >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	name >> g:GOSt >> g:Sorter >> g:Orth >> g:Cocoa >> Copy values	description	namespace
1	TP53	1.1	P04637	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, DBASS5, DBASS3, HGNC, WIKIGENE
2	MDM2	2.1	Q00987	MDM2	MDM2 proto-oncogene, E3 ubiquitin protein ligase [Source:HGNC Symbol;Acc:HGNC:6973]	UNIPROT_GN, ENTREZGENE, VEGA_GENE, HGNC, WIKIGENE
3	207105_S_AT	3.1	O00459	PIK3R2	phosphoinositide-3-kinase, regulatory subunit 2 (beta) [Source:HGNC Symbol;Acc:HGNC:8980]	AFFY_HG_U133_PLUS_2, AFFY_HG_FOCUS, AFFY_HG_U133A_2, AFFY_HG_U133A
4	P60484	4.1	P60484	PTEN	phosphatase and tensin homolog [Source:HGNC Symbol;Acc:HGNC:9588]	UNIPROTSWISSPROT



- g:Convert
  - http://biit.cs.ut.ee/gprofiler/gconvert.cgi
- Ensembl Biomart
  - http://www.ensembl.org

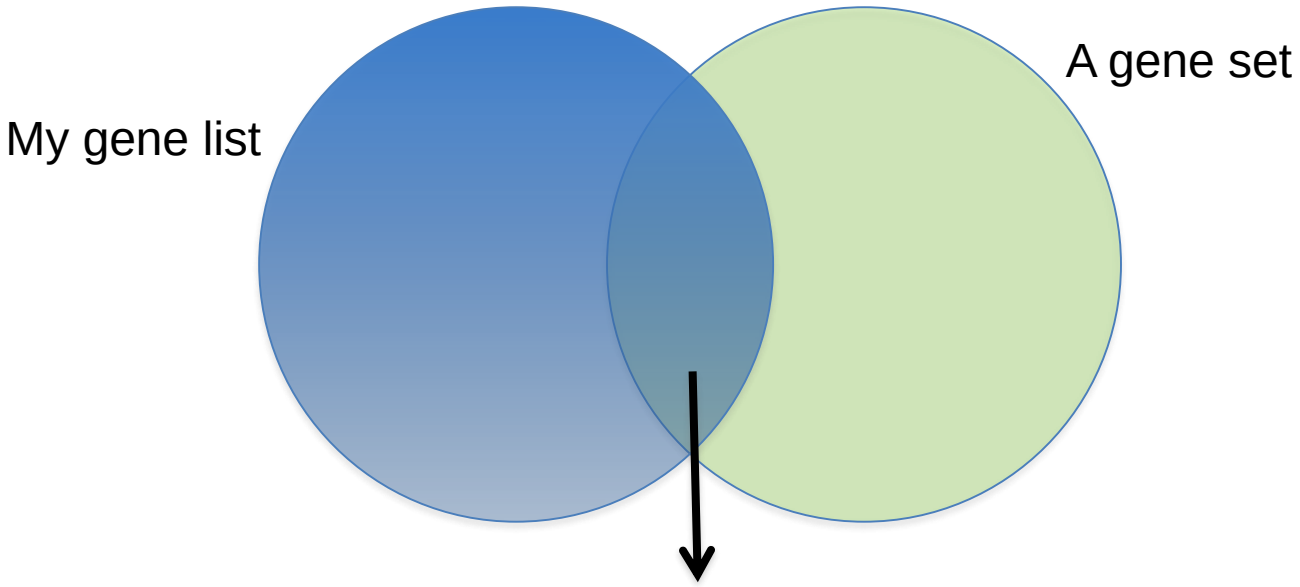
ID Challenges

- Avoid errors: map IDs correctly
  - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol: TP53
- Excel error-introduction
  - OCT4 is changed to October-4 (paste as text)
- Problems reaching 100% coverage
  - E.g. due to version issues
  - Use multiple sources to increase coverage

Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics BMC Bioinformatics. 2004 Jun 23;5:80

2.3. Gene Set Analysis --ORA

Over-representation analysis (ORA) is the task of identifying the pathways that contain a number of genes from our gene list that would be hard to find by chance alone.

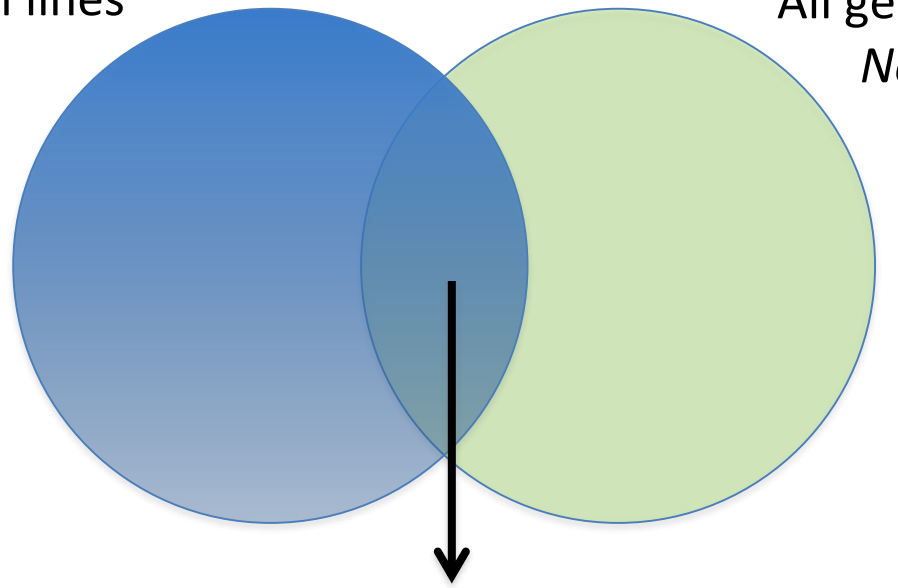


Are the genes in the intersection too many? What do we mean when we say "too many"? 5 out of 10? 7 out of 10? (We must use Statistics!)

Hypothesis: drug sensitivity in brain cancer is related to reduced neurotransmitter signaling

Gene list from experiment:
Genes down-regulated in drug-sensitive brain cancer cell lines

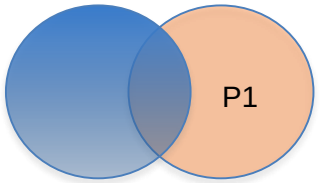
Pathway information:
All genes in the pathway called *Neurotransmitter signaling*



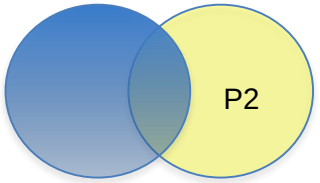
Statistical test: Are there more annotations in the gene list than expected by chance alone? ($p < 0.05$?)

Adapted from: Canadian Bioinformatics Workshop

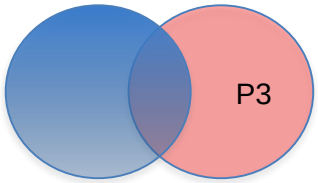
Usually, we do this for all gene sets in the database, and build a table



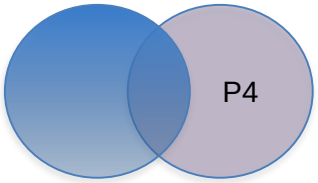
p-value = 0.04



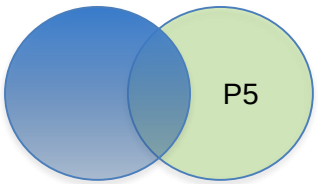
p-value = 0.2



p-value = 0.06



p-value = 0.003



p-value = 0.01

Gene Set	p-value
P4	0.003
P5	0.01
P1	0.04
P3	0.06
P2	0.2

Significant

Cutoff

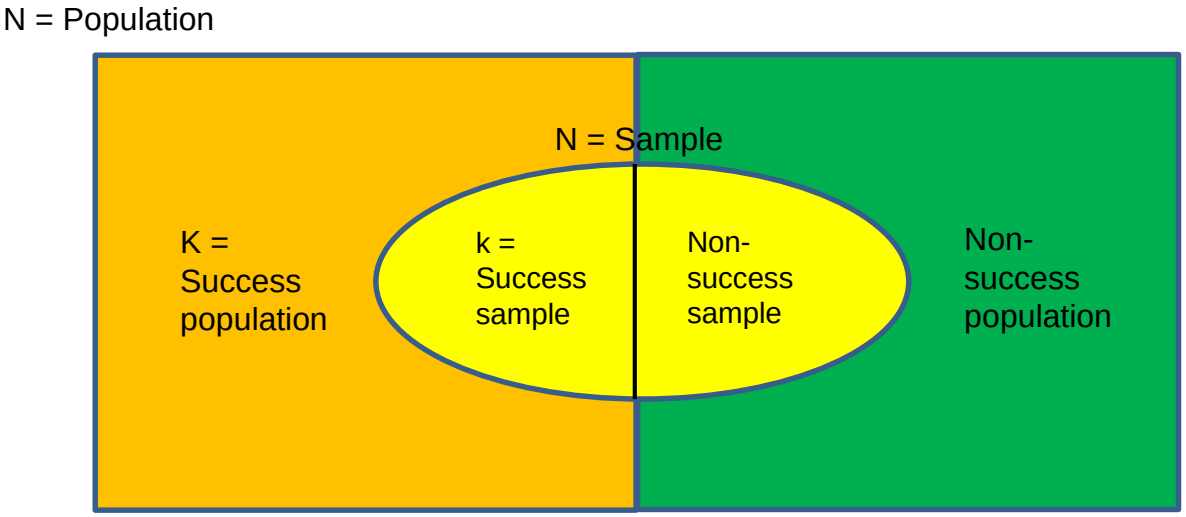
Statistical (Enrichment) Test:

What do you mean "enriched"? How many genes are “too many”?

The statistical formulation: If we randomly choose “n” genes, how likely is that all the “n” genes will be in a certain pathway?

If it is very unlikely (low probability), we say that the sample genes are over-represented in that pathway.

The most common ORA test is using the “Hypergeometric distribution” (HG).



N = Number of items in the population

K = Number of items in the population that we call “successes”

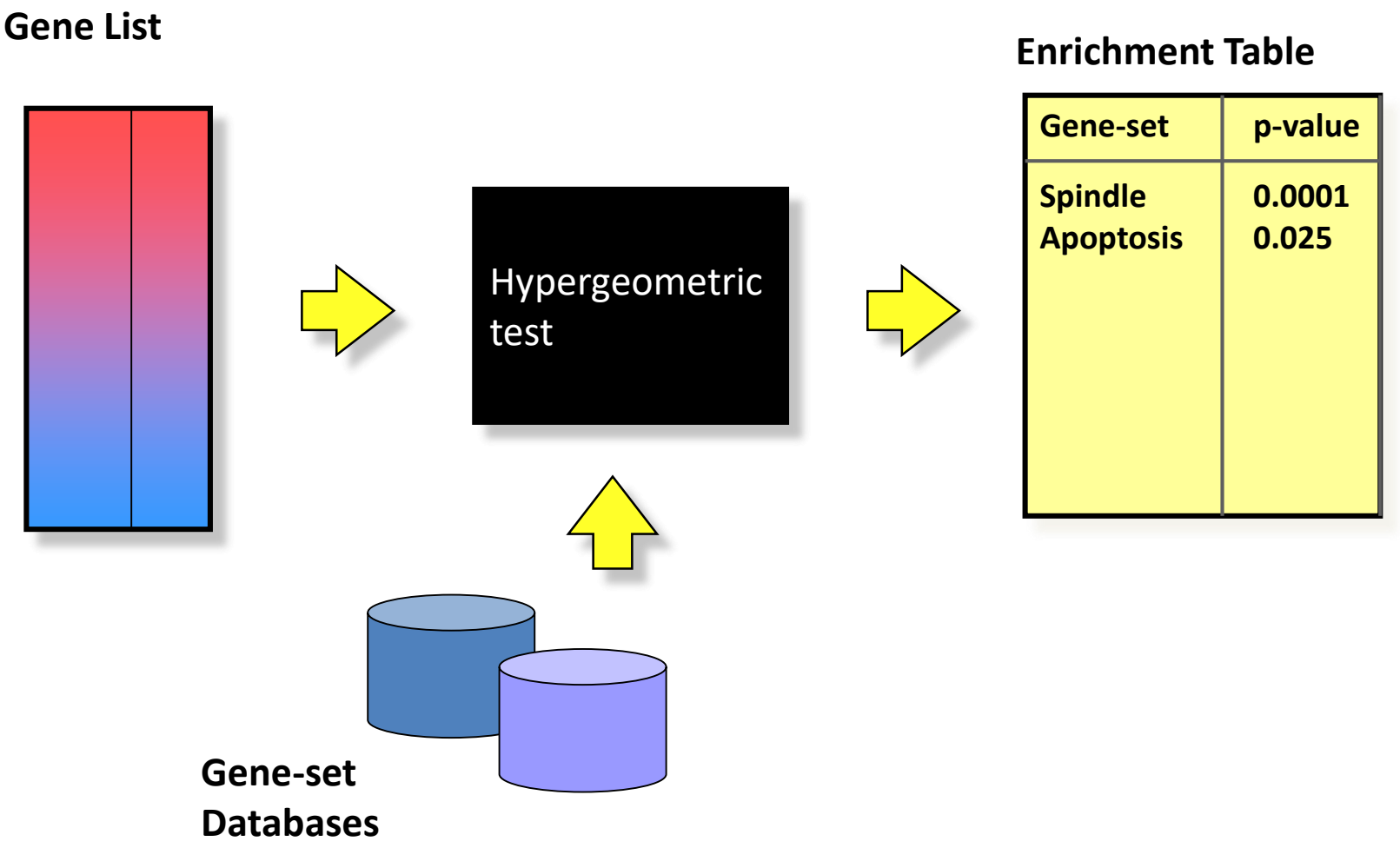
n = Number of items in the sample

k = Number of successes in the sample

Question: What is the probability of success P?

The HG describes the probability (P) of k successes in n draws, without replacement, from a population of size N that contains K successes.

The Statistical Test: Is this more enriched than expected by chance alone? Is it better than P?



Adapted from: Canadian Bioinformatics Workshop

Probability of success: P(X=k)

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} \quad \text{for } 0 \le k \le n,$$

$$\begin{aligned} n! &= \prod_{k=1}^{n} k \\ &= 1 \cdot 2 \cdot 3 \cdots (n-2) \cdot (n-1) \cdot n \\ &= n(n-1)(n-2) \cdots (2)(1) \end{aligned}$$

$$4! = 4 * 3 * 2 * 1$$

Example: Suppose we randomly select 5 cards without replacement from a deck of cards. What is the probability of getting exactly 2 red cards?

N = Population = All cards in the deck = 52

K = Population success = All red cards in the deck = 26

n = Sample = 5

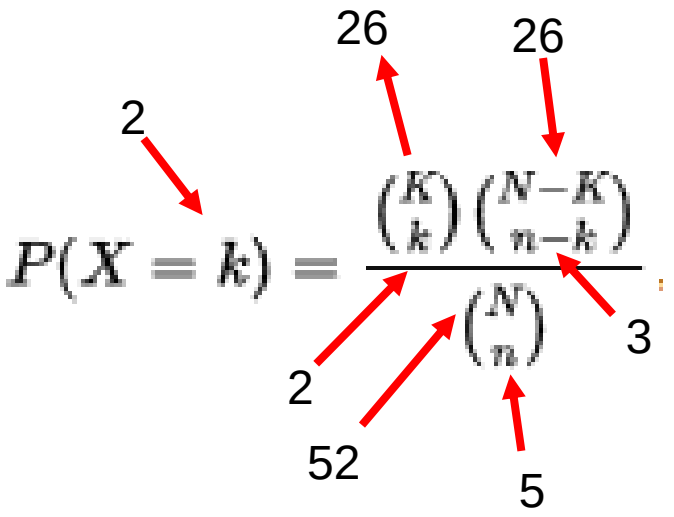
k = Sample success = 2

N – K = 26

n – k = 3

What is the probability of success?

Probability of success: P(X=k)



$$P(X=2)=\frac{\binom{26}{2}\binom{26}{3}}{\binom{52}{5}}$$

$$P(X=2)=\frac{325*2600}{2598960}=0.3251$$

Example: We have 52 students, 26 tall and 26 small. Suppose we randomly select 5 students from the group. What is the probability of getting exactly 2 tall students?

N = Population = All students = 52

K = Population success = All tall students = 26

n = Sample = 5

k = Sample success = Tall students in the sample = 2

N – K = 26

n – k = 3

What is the probability of success?

Probability of success: P(X=k)

$$P(X=2)=\frac{\binom{26}{2}\binom{26}{3}}{\binom{52}{5}}$$

$$P(X=2)=\frac{325*2600}{2598960}=0.3251$$

Example: Suppose we are using a database with 52 genes distributed in two pathways, each having 26 genes. Suppose we found 5 differentially-expressed genes in our experiment. What is the probability of getting exactly 2 genes in pathway A?

N = Population = All genes in the database = 52

K = Population success = All genes in pathway A = 26

n = Sample = Our full set of DEG = 5

k = Sample success = 2

N – K = 26

n – k = 3

Probability of success: P(X=k)

$$P(X=2)=\frac{\binom{26}{2}\binom{26}{3}}{\binom{52}{5}}$$

$$P(X=2)=\frac{325*2600}{2598960}=0.3251$$

- **But our original question was not the probability of success. The question was if the genes are enriched (over-represented) in that pathway or not.**

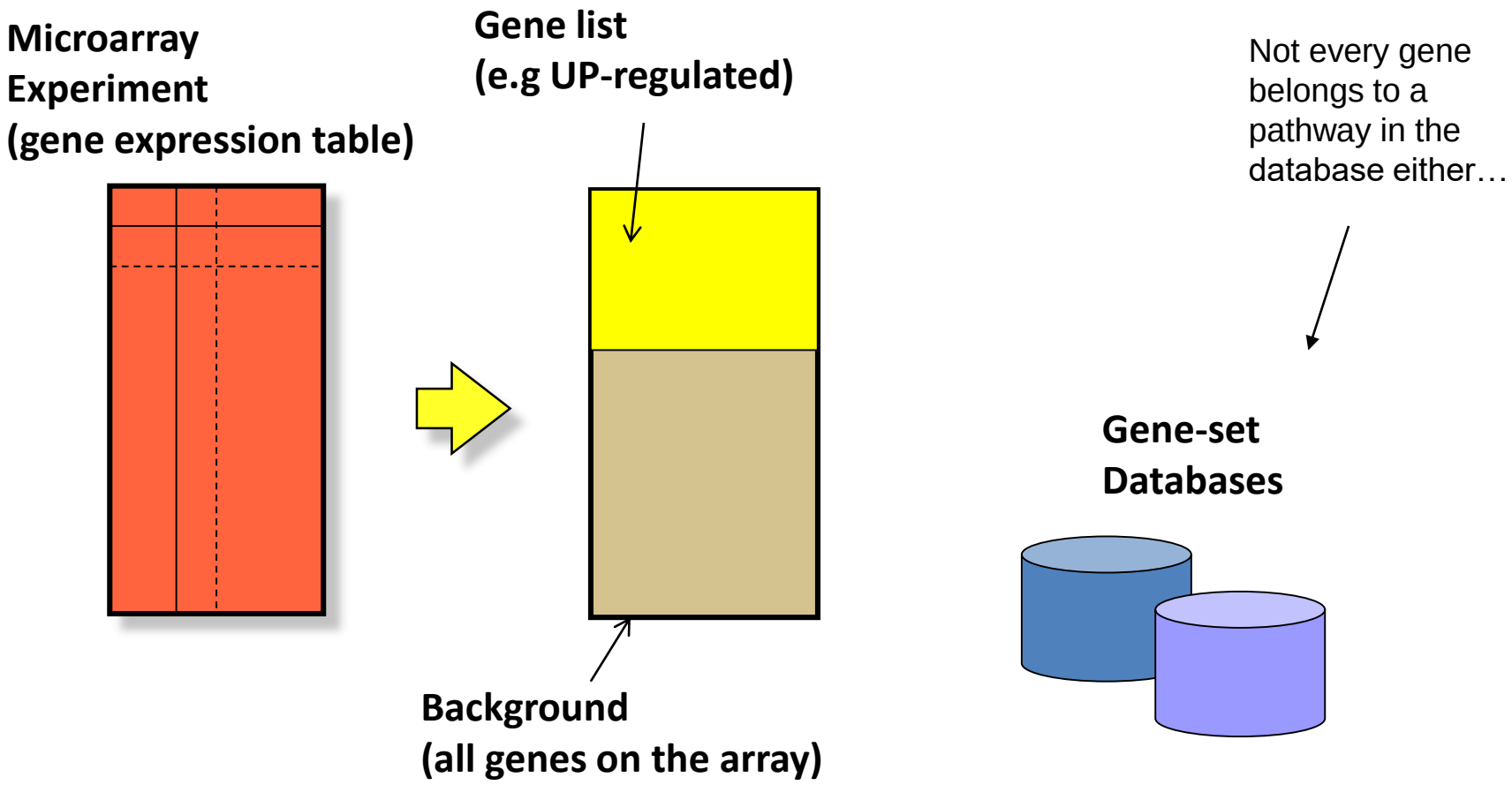
- We usually accept a threshold of p = 0.05 to decide that.

- Our p = 0.3251 is much higher than that, which means that is easy for those two genes to appear in pathway A just by chance. Therefore, we say that those two genes are not enriched in pathway A.

- ORA tools search for over-representation in a given database of pathways.
- In each case, the sample success is the intersection between our list of genes and one specific pathway (f.ex., if there are 3 genes of our list in pathway B, k=3 for pathway B).
- The tool shows as results the pathways with p smaller than our threshold (usually, 0.05).

The Background

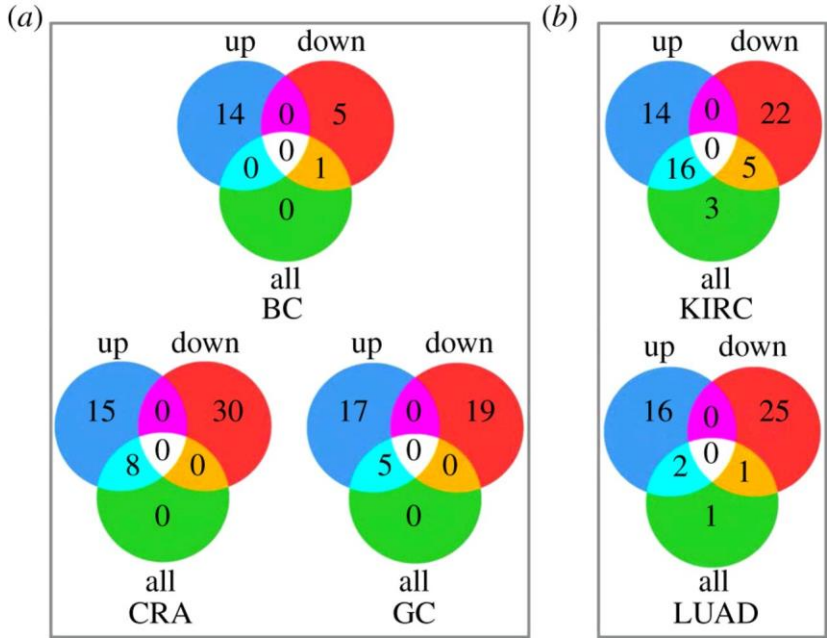
Need to choose “background population” appropriately, e.g., if only portion of the total gene complement is queried (or available for annotation), only use that population as background.



Adapted from: Canadian Bioinformatics Workshop

Should we analyze all genes together? Or separate analyses for up-regulated and down-regulated?

five types of tumours, we illustrate that the separate analysis of up- and down-regulated genes could identify more pathways that are really pertinent to phenotypic difference. In conclusion, analysing up- and downregulated genes separately is more powerful than analysing all of the DE genes together.



HONG G., Separate enrichment analysis of pathways for up- and downregulated genes, 2013

Should we use all genes in a pathway or gene set?

Some authors filter the gene sets:

Remove gene sets with only a few genes and those with a very large number of genes.

Some authors prefer to divide large pathways into sub-pathways:

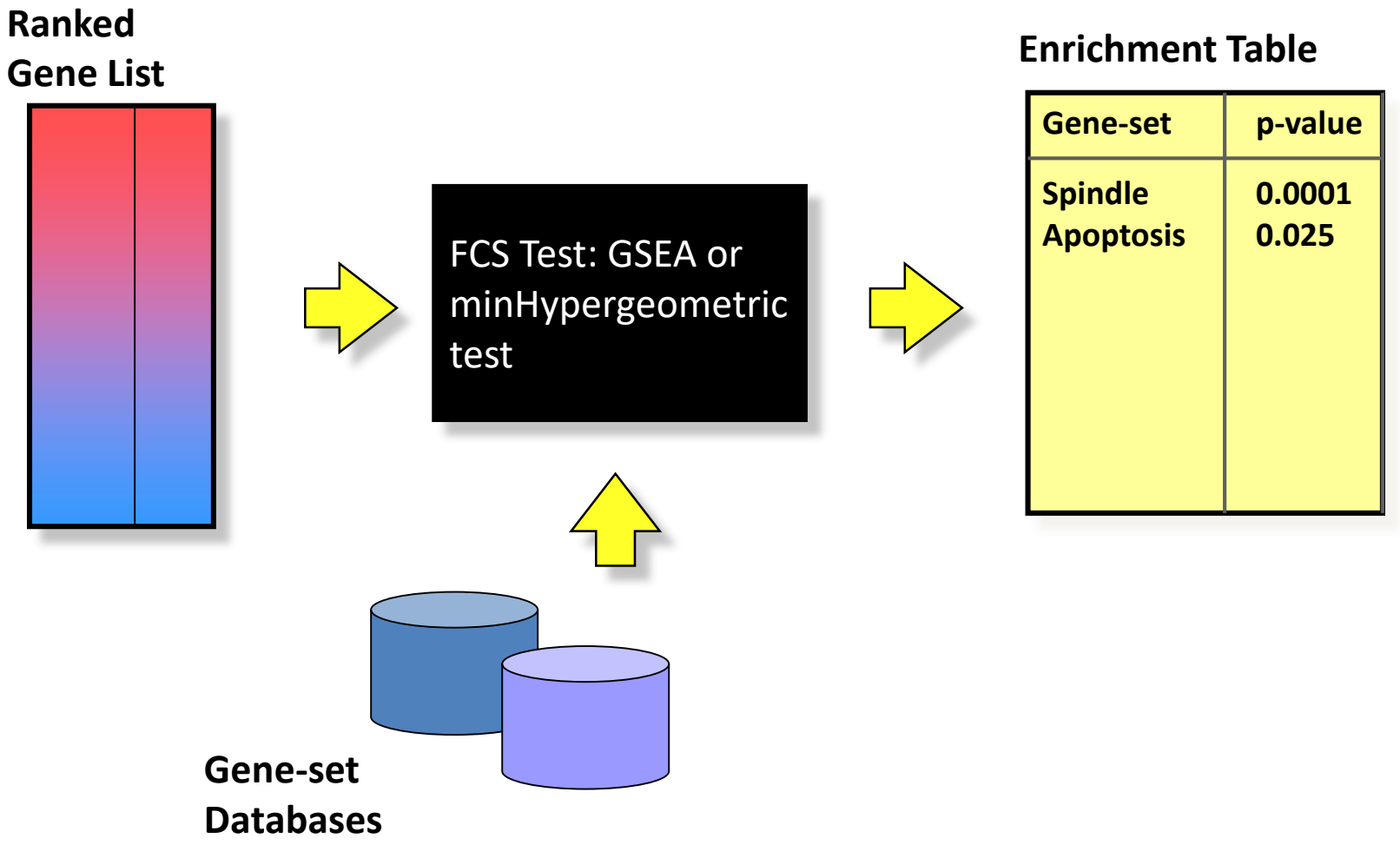
Low et al. [67] divided the estrogen metabolic pathway into three sub-pathways involved in androgen synthesis, androgen-to-estrogen conversion and estrogen removal and then found only SNPs within the androgen-to-estrogen conversion pathway were significantly associated with breast and endometrial cancer susceptibilities.

2.4. Gene Set Analysis --FCS

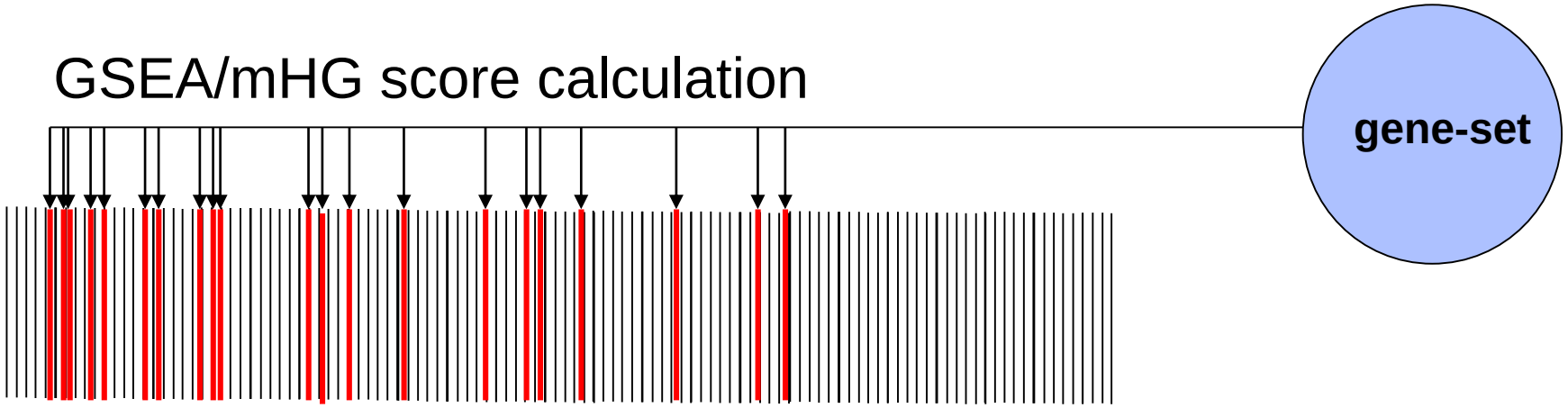
Problems with gene lists

- Threshold for up- and down-regulated genes is arbitrary (f.ex., fold-change > 2, or log-fold-change > 1.5)
- We get different results at different threshold settings.
- Changes in pathway activity can happen not only if we have a few highly differentially expressed genes but also if we have multiple genes more modestly differentially expressed.

Functional Class Scoring (FCS)



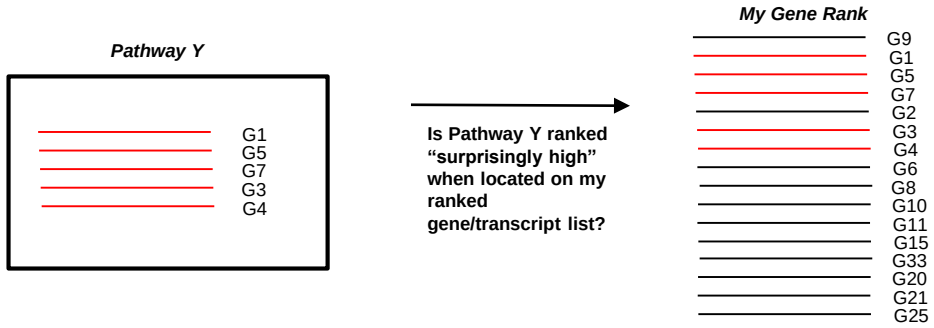
How to score a gene set?



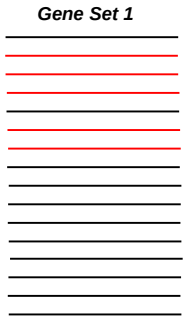
Where are the gene-set genes located in the ranked list?
Is there distribution random, or is there an enrichment in either end?

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

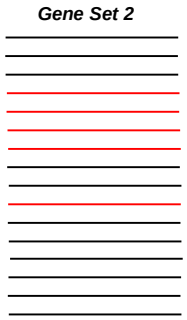
How to score a gene set?



Scoring a gene set using the mean rank:



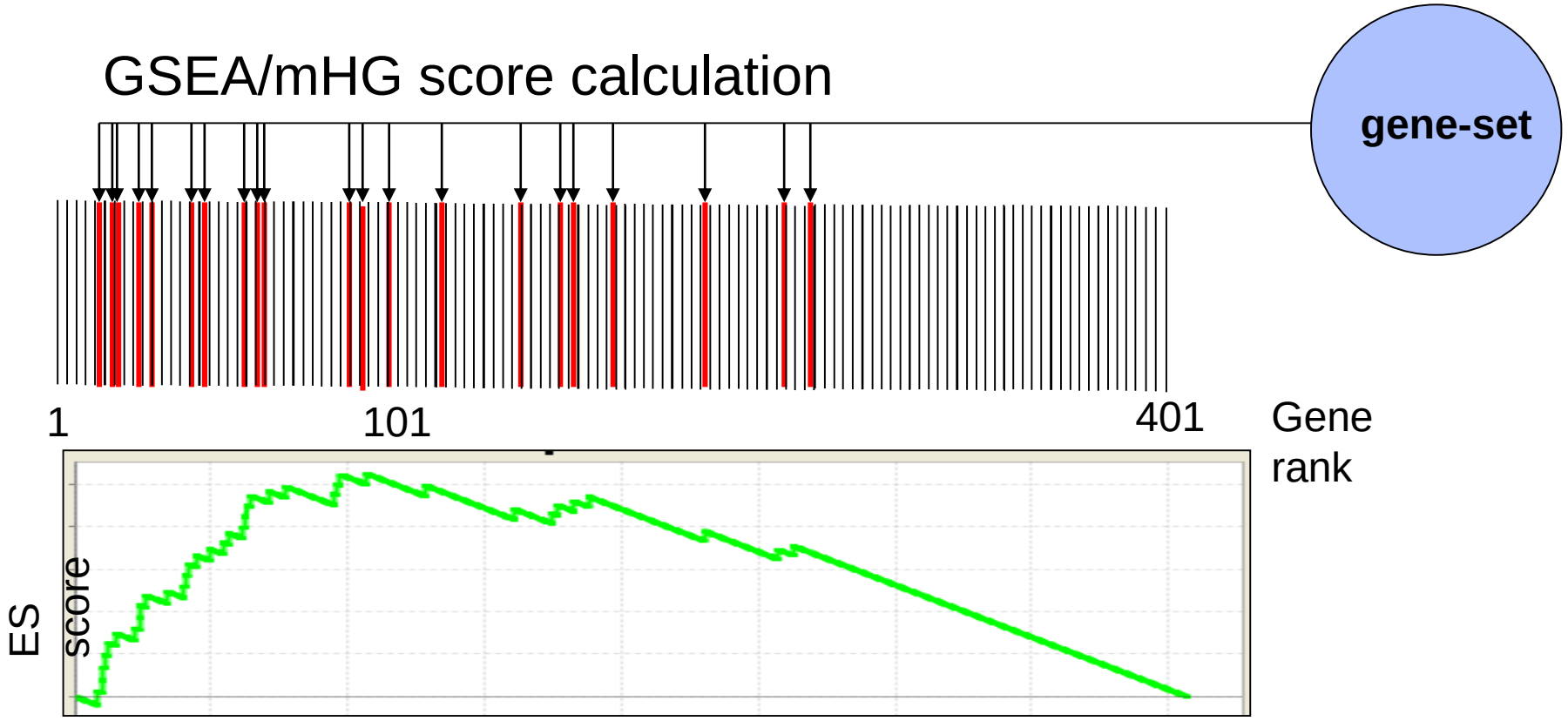
Mean Rank = (2+3+4+6+7) / 5 = 4.4



Mean Rank = (4+5+6+7+10) / 5 = 6.4

There are more complex scoring methods, such as: KS, max-mean, and others

GSEA/mHG: Method

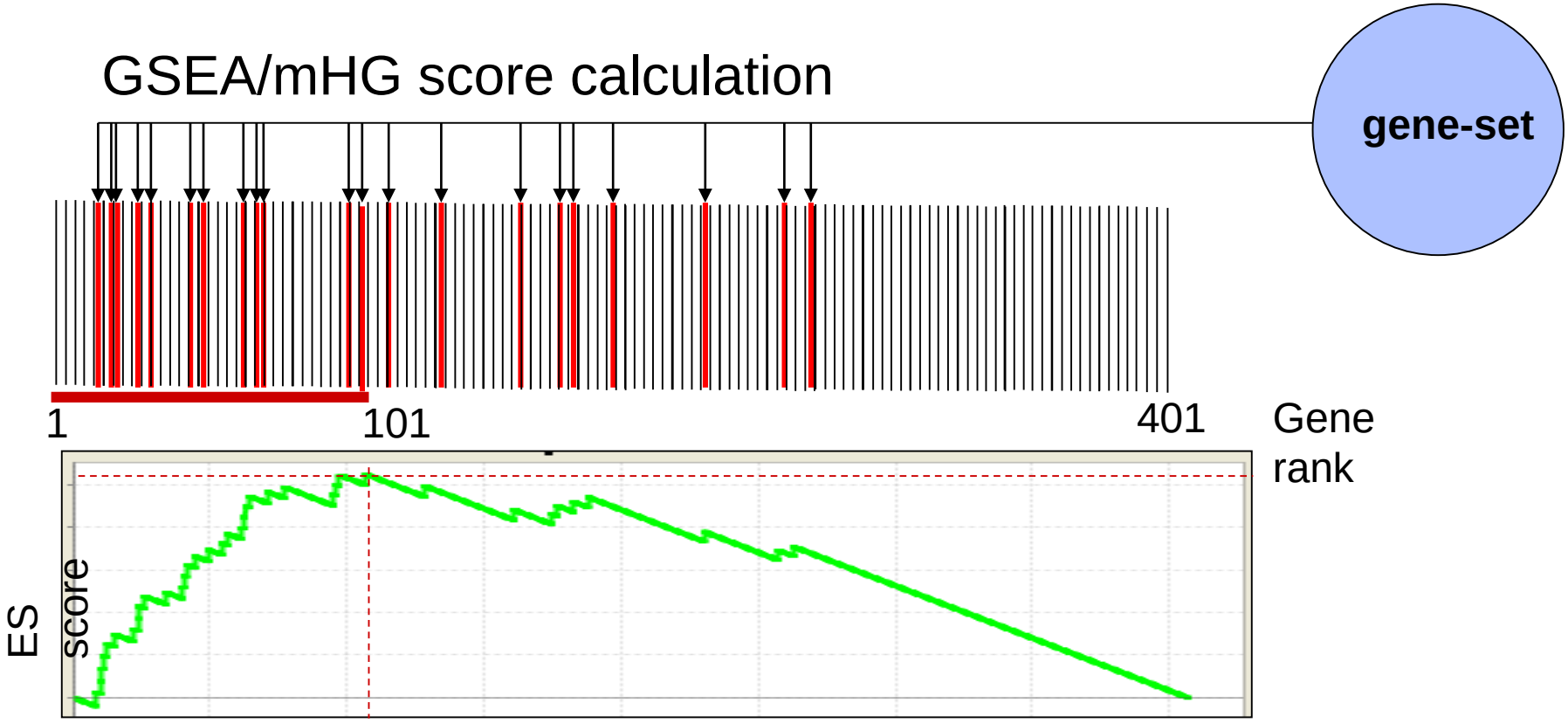


Every present gene (thick red vertical bar) gives a positive contribution,
Every absent gene (black vertical bar) gives a negative contribution

For mHG, ES score = -log P of hypergeometric test at that threshold

Warning: the alignment here between bars and plot is a little off

GSEA/mHG: Method



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.

Going from ES score to p-value

We can compute an empirical p-value using permutations, in the following way:

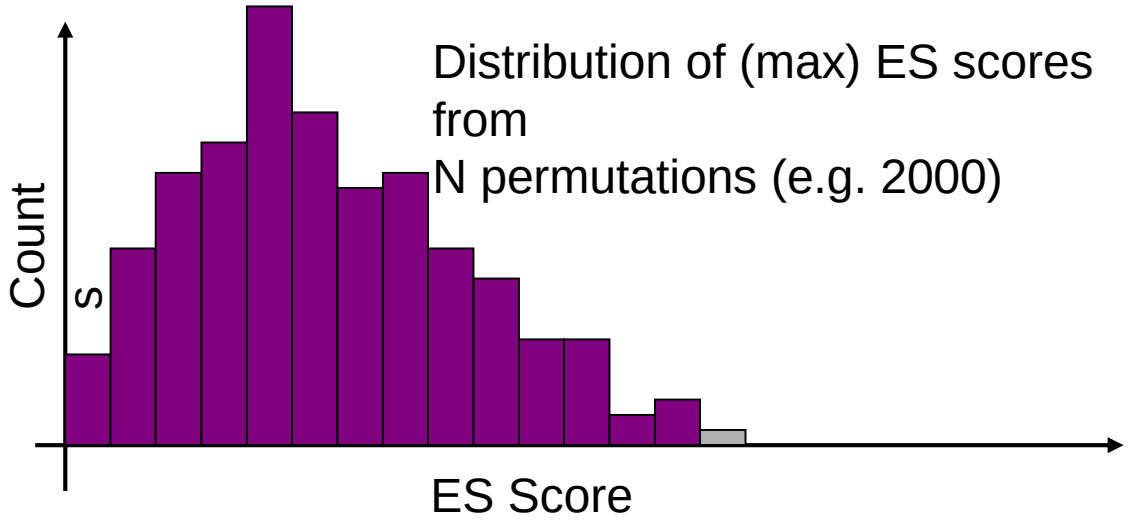
1. Transforming the gene rank into “n” random ranks and then applying the previous procedure in each case. In the end, we will end up with “n” ES values from the random cases.

2. Then we will compare our real ES value to all the “n” random ones. Ideally, our ES value should be higher than the random ones, but it is possible to get some cases where it is smaller just by chance. The ratio of times that a random ES is better than the real one, is our p-value. 5 successes of the random ES out of 100 trials would mean a p-value of 0.05.

In statistical terms…

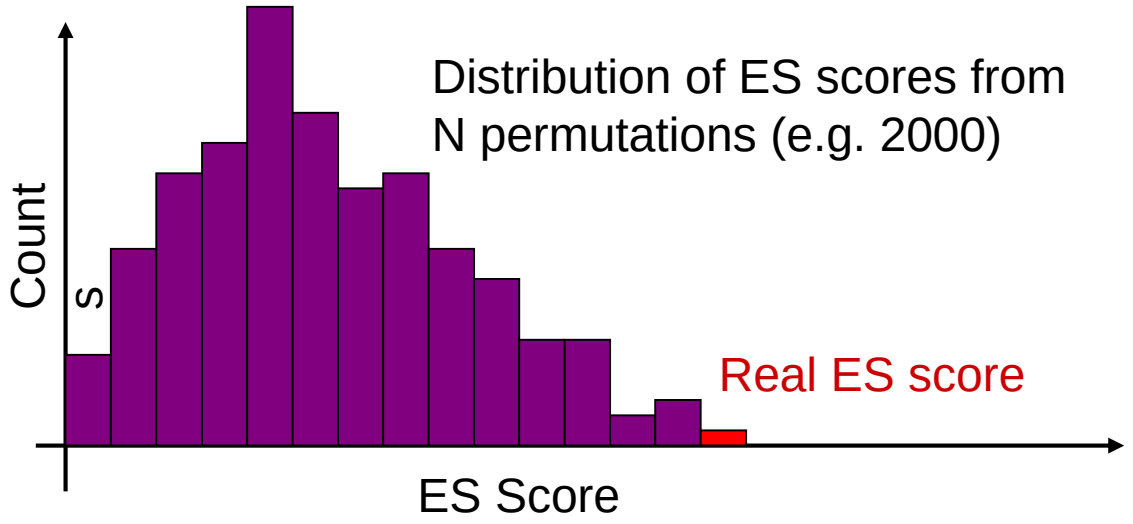
Empirical p-value estimation (for every gene-set)

1. Generate null-hypothesis distribution from randomized data



In statistical terms…

Estimate empirical p-value by comparing observed max 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

2.5. Multiple testing correction

Multiple testing correction

A $p<0.05$ means that there is still a 5% probability of finding some correlation purely by chance. This is a small number, but if you play it 1000 times, it gets very probable that you will find a positive result just by chance.

Therefore, a ***correction for multiple testing*** is needed. Some of the methods include ***Bonferroni*** and ***False Discovery Rate (FDR)***.

Simple P-value correction: Bonferroni

* If M = # Tests:

Corrected p-value = M * original p-value

- In other words, we are looking for $p<0.05/M$. If M is 1000 tests (1000 pathways, f.ex.), now p must be less than 0.00005
- Bonferroni correction is very stringent and can “wash away” real enrichments leading to false negatives

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 I

Rank	Category	(Nominal) P-value
1	*Transcriptional*	0.001
2	*regulation*	0.002
3	*Transcription factor*	0.003
4	*Initiation of transcription*	0.0031
5	*Nuclear localization*	0.005
…	*Chromatin modification*	…
	…	
52		0.97
53	*Cytoplasmic localization Translation*	0.99

Sort P-values of all tests in increasing order

Benjamini-Hochberg example II

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

Adjusted P-value is "nominal" P-value times # of tests divided by the rank of the P-value in sorted list

Adjusted P-value = P-value X [# of tests] / Rank

Benjamini-Hochberg example III

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

Benjamini-Hochberg example III

P-value threshold for FDR < 0.05

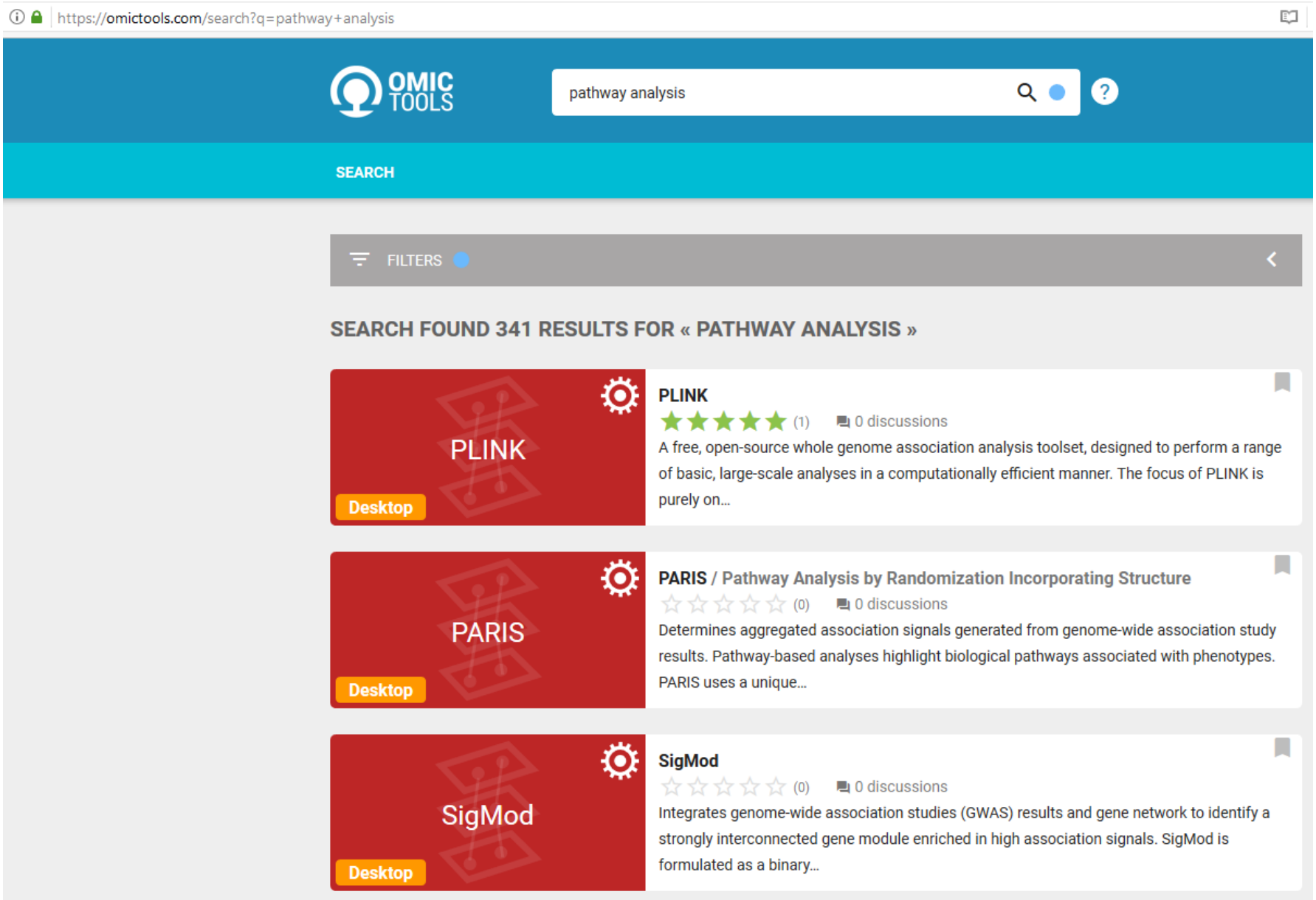
Rank	Category	(Nominal) P-value	Adjusted P-value	FDR / Q-value
1	*Transcriptional*	0.001	0.001 x 53/1 = 0.053	0.040
2	*regulation*	0.002	0.002 x 53/2 = 0.053	0.040
3	*Transcription factor*	0.003	0.003 x 53/3 = 0.053	0.040
4	*Initiation of transcription*	**0.0031**	0.0031 x 53/4 = 0.040	0.040
5	*Nuclear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

Red: non-significant
Green: significant at FDR < 0.05

P-value threshold is highest ranking P-value for which corresponding Q-value is below desired significance threshold

2.5. Gene Set Analysis --Software

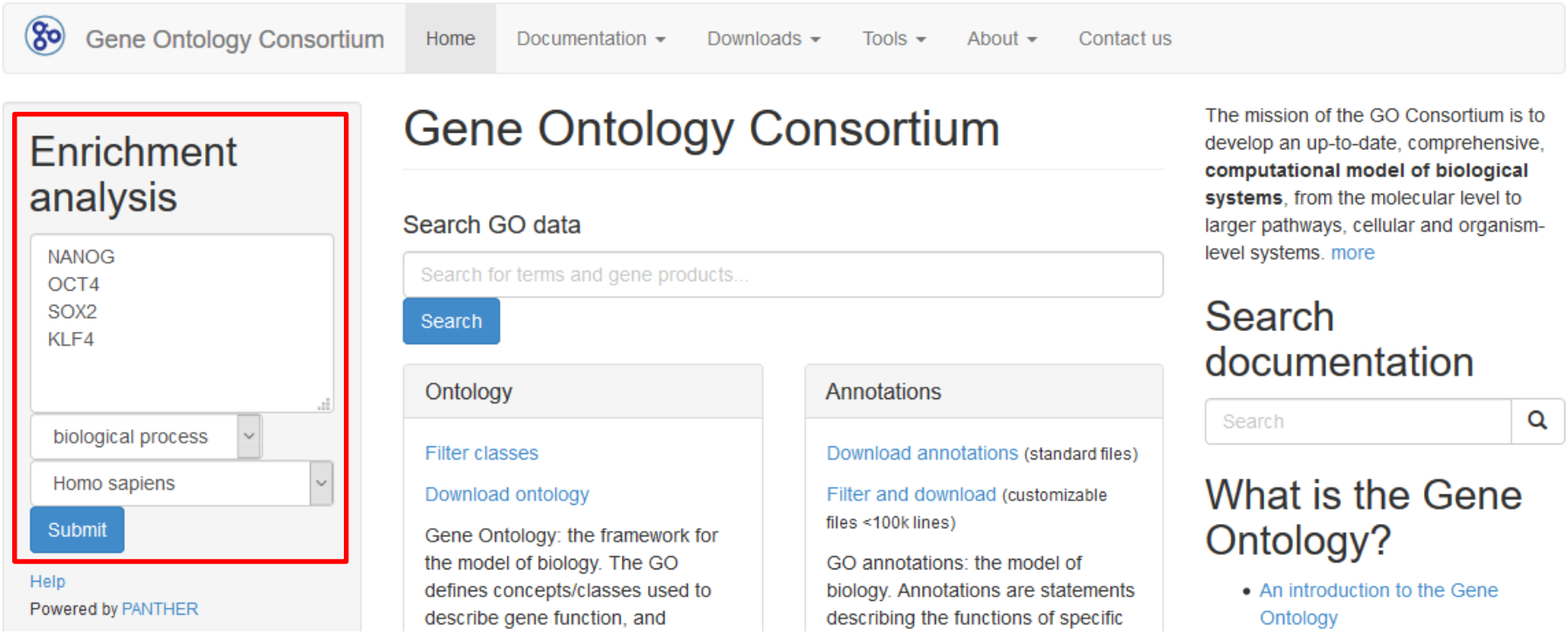
Where to find software?: Omicstools



How to learn to use new software?

1. Try to find tutorials (or “vignettes” in R).
2. Read the manuals to see all other options that were not covered in the tutorials.
3. Ask questions. Don’t be afraid to ask (but ask after you tried first).

GO


Gene Ontology Consortium
Home
Documentation
Downloads
Tools
About
Contact us
Enrichment analysis
NANOG
OCT4
SOX2
KLF4
biological process
Homo sapiens
Submit
Help
Powered by PANTHER
Gene Ontology Consortium
Search GO data
Search for terms and gene products...
Search
Ontology
Filter classes
Download ontology
Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and
Annotations
Download annotations (standard files)
Filter and download (customizable files <100k lines)
GO annotations: the model of biology. Annotations are statements describing the functions of specific
The mission of the GO Consortium is to develop an up-to-date, comprehensive, computational model of biological systems, from the molecular level to larger pathways, cellular and organism-level systems. more
Search documentation
Search
What is the Gene Ontology?
An introduction to the Gene Ontology

GO

	Homo sapiens (REF)	upload 1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
endodermal cell fate specification	6	2	.00	> 100	+	3.79E-07	9.95E-04
↳endodermal cell fate commitment	12	2	.00	> 100	+	1.23E-06	2.42E-03
↳endodermal cell differentiation	40	2	.01	> 100	+	1.17E-05	1.83E-02
↳endoderm formation	46	2	.01	> 100	+	1.53E-05	2.18E-02
↳endoderm development	72	2	.01	> 100	+	3.65E-05	4.42E-02
↳formation of primary germ layer	106	3	.02	> 100	+	1.35E-07	7.09E-04
↳gastrulation	152	3	.02	> 100	+	3.92E-07	8.81E-04
↳embryonic morphogenesis	556	3	.08	37.85	+	1.86E-05	2.44E-02
↳cell fate commitment involved in formation of primary germ layer	26	3	.00	> 100	+	2.35E-09	3.70E-05
↳cell fate commitment	232	3	.03	90.70	+	1.37E-06	2.40E-03
↳cell fate specification	73	2	.01	> 100	+	3.75E-05	4.22E-02
somatic stem cell population maintenance	53	3	.01	> 100	+	1.78E-08	1.40E-04
↳stem cell population maintenance	124	3	.02	> 100	+	2.15E-07	8.44E-04
↳maintenance of cell number	127	3	.02	> 100	+	2.30E-07	7.25E-04

Pathway enrichment analysis software: DAVID

Pathway enrichment analysis software: DAVID

Upload | List | Background

Upload Gene List

Demolist 1 Demolist 2

Upload Help

Step 1: Enter Gene List

A: Paste a list

```
79923
5460
6657
9314
```

Clear

Or

B:Choose From a File

Browse... No file selected.

☐ Multi-List File

Step 2: Select Identifier

ENTREZ_GENE_ID

Step 3: List Type

Gene List ◉

Background ○

Step 4: Submit List

Submit List

Gene List Report

Help and Manual

Current Gene List: List_3

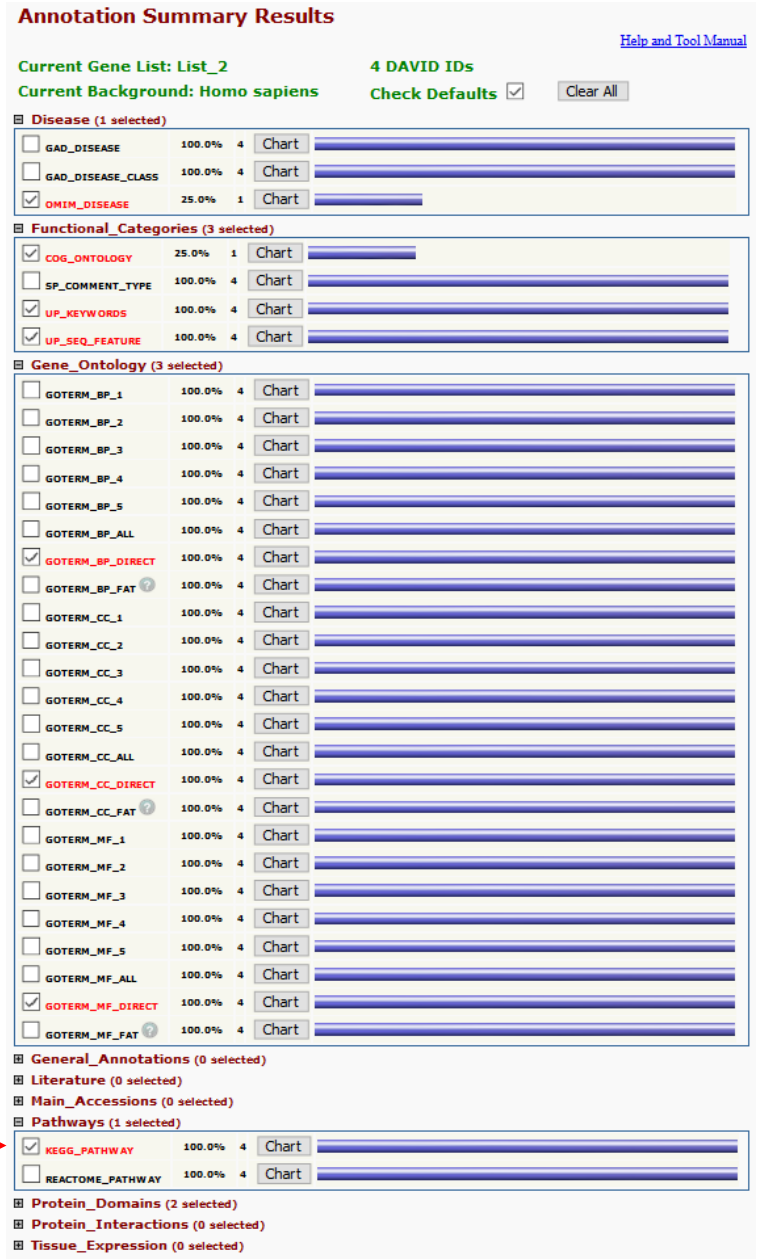
Current Background: Homo sapiens

4 DAVID IDs

Download File

ENTREZ_GENE_ID	Gene Name	Related Genes	Species
5460	POU class 5 homeobox 1(POU5F1)	RG	Homo sapiens
6657	SRY-box 2(SOX2)	RG	Homo sapiens
79923	Nanog homeobox(NANOG)	RG	Homo sapiens
9314	Kruppel like factor 4(KLF4)	RG	Homo sapiens

Pathway enrichment analysis software: DAVID



Results for KEGG Pathways

Pathway enrichment analysis software: DAVID

Functional Annotation Clustering

Help and Manual

Current Gene List: List_2
Current Background: Homo sapiens
4 DAVID IDs

⊞ Options Classification Stringency: Medium

Rerun using options | Create Sublist

1 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 2.4	G		Count	P_Value	Benjamini
☐	GOTERM_BP_DIRECT	somatic stem cell population maintenance	RT		4	5.5E-8	4.8E-6
☐	GOTERM_BP_DIRECT	endodermal cell fate specification	RT		3	2.1E-7	9.1E-6
☐	KEGG_PATHWAY	Signaling pathways regulating pluripotency of stem cells	RT		4	8.1E-6	2.4E-5
☐	GOTERM_BP_DIRECT	regulation of gene expression	RT		3	1.0E-4	3.0E-3
☐	GOTERM_MF_DIRECT	transcription factor activity, sequence-specific DNA binding	RT		4	1.8E-4	5.1E-3
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		4	2.0E-4	4.3E-3
☐	GOTERM_MF_DIRECT	transcription regulatory region DNA binding	RT		3	4.7E-4	6.6E-3
☐	UP_KEYWORDS	DNA-binding	RT		4	9.9E-4	2.0E-2
☐	UP_KEYWORDS	Transcription regulation	RT		4	1.5E-3	1.4E-2
☐	UP_KEYWORDS	Transcription	RT		4	1.6E-3	1.0E-2
☐	GOTERM_BP_DIRECT	transcription from RNA polymerase II promoter	RT		3	2.7E-3	4.6E-2
☐	GOTERM_MF_DIRECT	sequence-specific DNA binding	RT		3	2.8E-3	2.5E-2
☐	UP_KEYWORDS	Activator	RT		3	3.0E-3	1.5E-2
☐	GOTERM_CC_DIRECT	nucleoplasm	RT		4	3.6E-3	3.2E-2
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		3	5.4E-3	7.4E-2
☐	UP_KEYWORDS	Developmental protein	RT		3	6.2E-3	2.4E-2
☐	UP_KEYWORDS	Isopeptide bond	RT		3	8.7E-3	2.9E-2
☐	UP_KEYWORDS	Nucleus	RT		4	1.7E-2	4.7E-2
☐	UP_KEYWORDS	Ubl conjugation	RT		3	1.9E-2	4.8E-2
☐	GOTERM_BP_DIRECT	regulation of transcription, DNA-templated	RT		3	2.3E-2	2.5E-1
☐	GOTERM_MF_DIRECT	DNA binding	RT		3	2.8E-2	1.8E-1

Enrichr

Input data

Choose an input file to upload. Either in BED format or a list of genes. For a quantitative set, add a comma and the level of membership of that gene. The membership level is a number between 0.0 and 1.0 to represent a weight for each gene, where the weight of 0.0 will completely discard the gene from the enrichment analysis and the weight of 1.0 is the maximum.

Try an example BED file.

Browse... No file selected.

Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try two examples: crisp set example, fuzzy set example

NANOG
OCT4
SOX2
KLF4

0 gene(s) entered

Enter a brief description for the list in case you want to share it. (Optional)

Submit

Contribute

Enrichr

Description No description available (4 genes)

KEGG 2016

- Signaling pathways regulating pluripotency
- Proteoglycans in cancer_Homo sapiens_hsa
- Hippo signaling pathway_Homo sapiens_hsa

WikiPathways 2016

- Preimplantation Embryo_Homo sapiens_WP
- Mesodermal Commitment Pathway_Homo s
- PluriNetWork_Mus musculus_WP1763
- Cardiac Progenitor Differentiation_Homo sa
- Endoderm Differentiation_Homo sapiens_WI

ARCHS4 Kinases Coexp

- ACVR2B_human_kinase_ARCHS4_coexpressi
- ROR1_human_kinase_ARCHS4_coexpression
- TAOK3_human_kinase_ARCHS4_coexpressio
- HUNK_human_kinase_ARCHS4_coexpressio
- PAN3_human_kinase_ARCHS4_coexpression

Reactome 2016

- Transcriptional regulation of pluripotent ste
- POU5F1 (OCT4), SOX2, NANOG activate gene
- POU5F1 (OCT4), SOX2, NANOG repress gene
- Developmental Biology_Homo sapiens_R-HS
- Synthesis, secretion, and deacylation of Ghr

BioCarta 2016

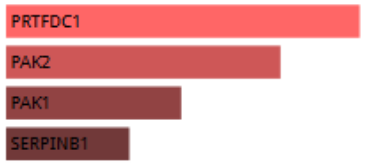
HumanCyc 2016

NCI-Nature 2016

- Regulation of nuclear beta catenin signaling

Panther 2016

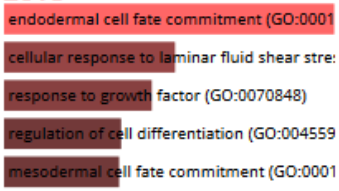
BioPlex 2017



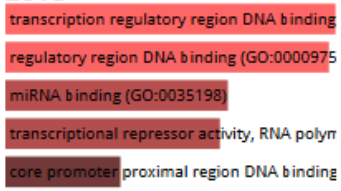
Enrichr

Description No description available (4 genes)

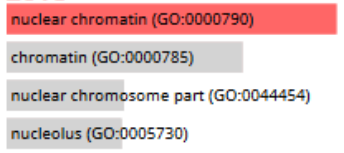
GO Biological Process 2018



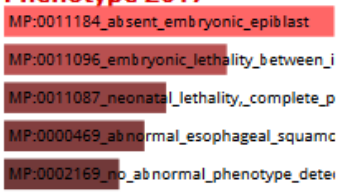
GO Molecular Function 2018



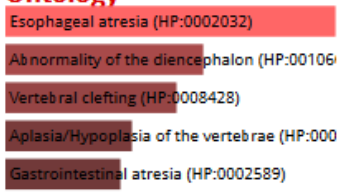
GO Cellular Component 2018



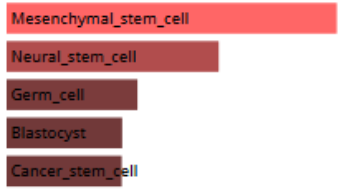
MGI Mammalian Phenotype 2017



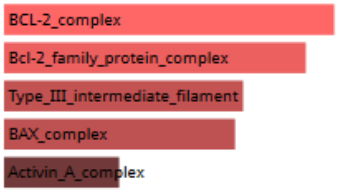
Human Phenotype Ontology



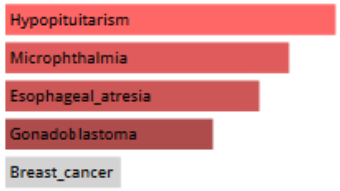
Jensen TISSUES



Jensen COMPARTMENTS



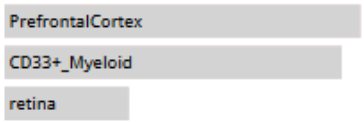
Jensen DISEASES



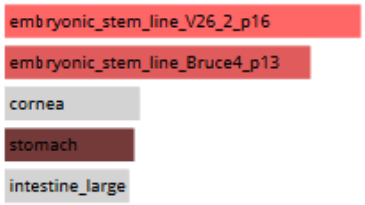
Enrichr

Description No description available (4 genes)

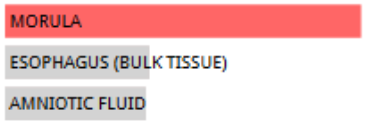
Human Gene Atlas



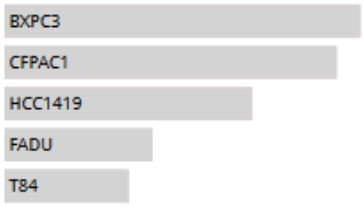
Mouse Gene Atlas



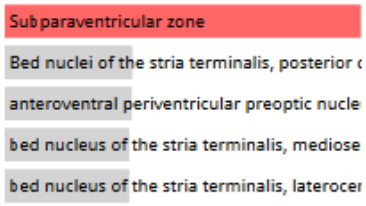
ARCHS4 Tissues



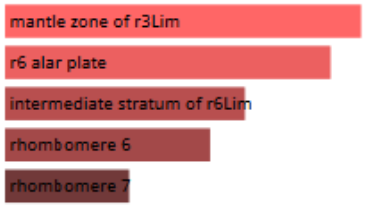
ARCHS4 Cell-lines



Allen Brain Atlas up

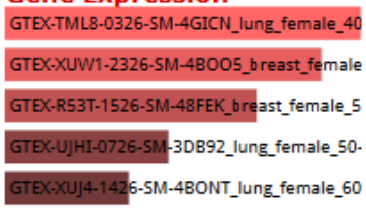


Allen Brain Atlas down

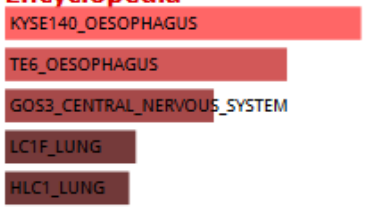


GTEx Tissue Sample Gene Expression

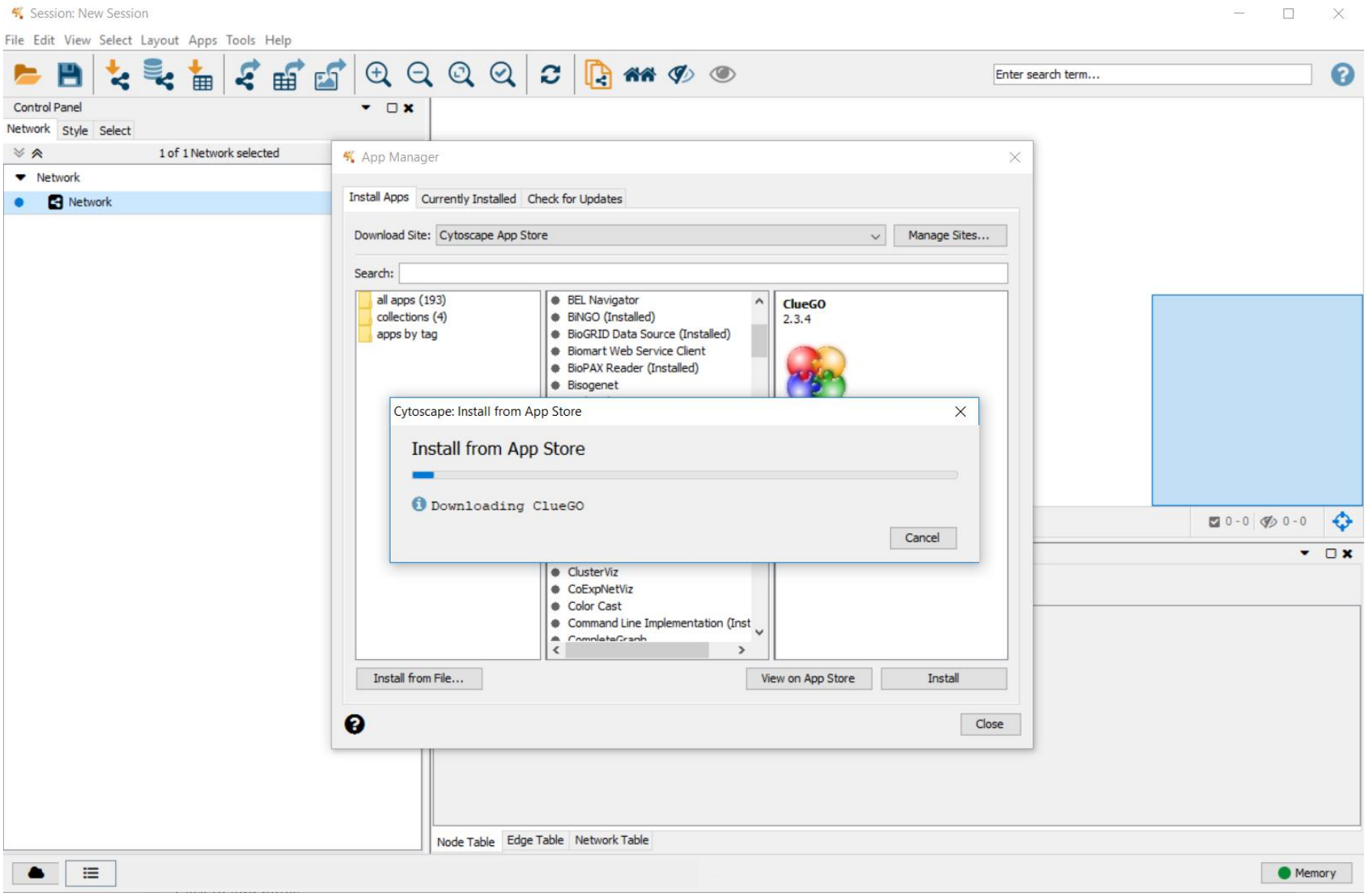
GTEx Tissue Sample Gene Expression



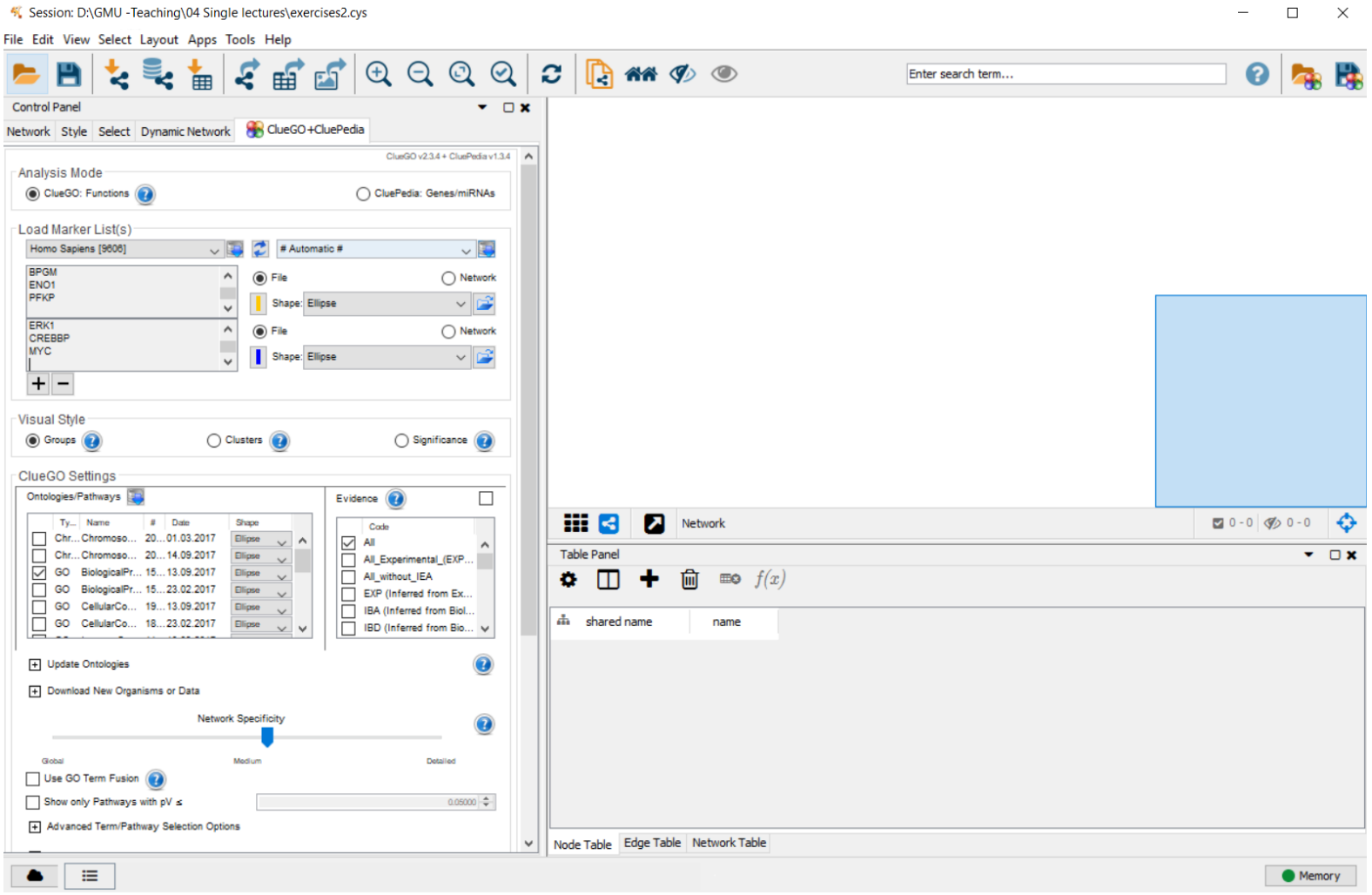
Cancer Cell Line Encyclopedia



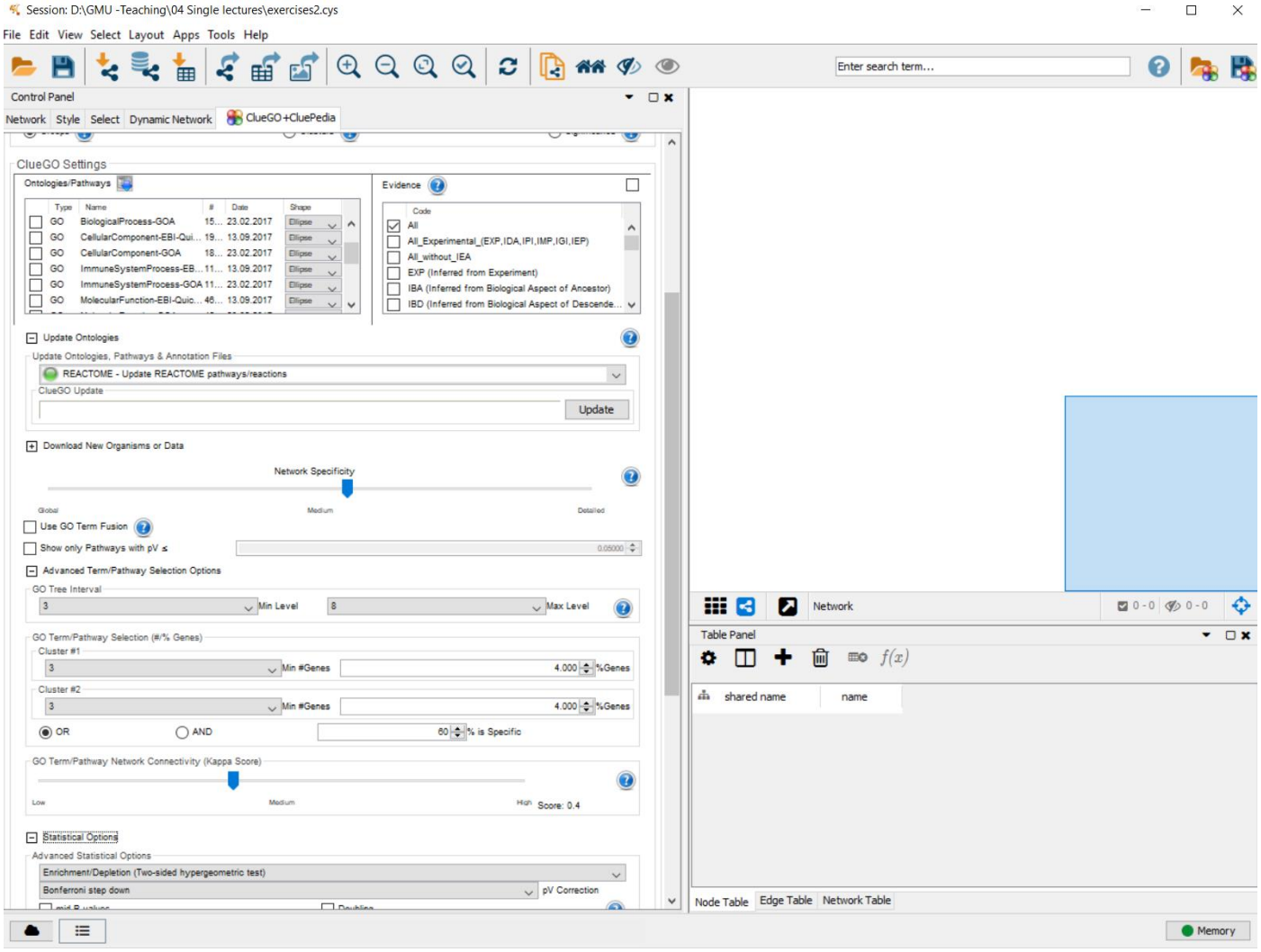
Pathway enrichment analysis software: Cytoscape / ClueGO



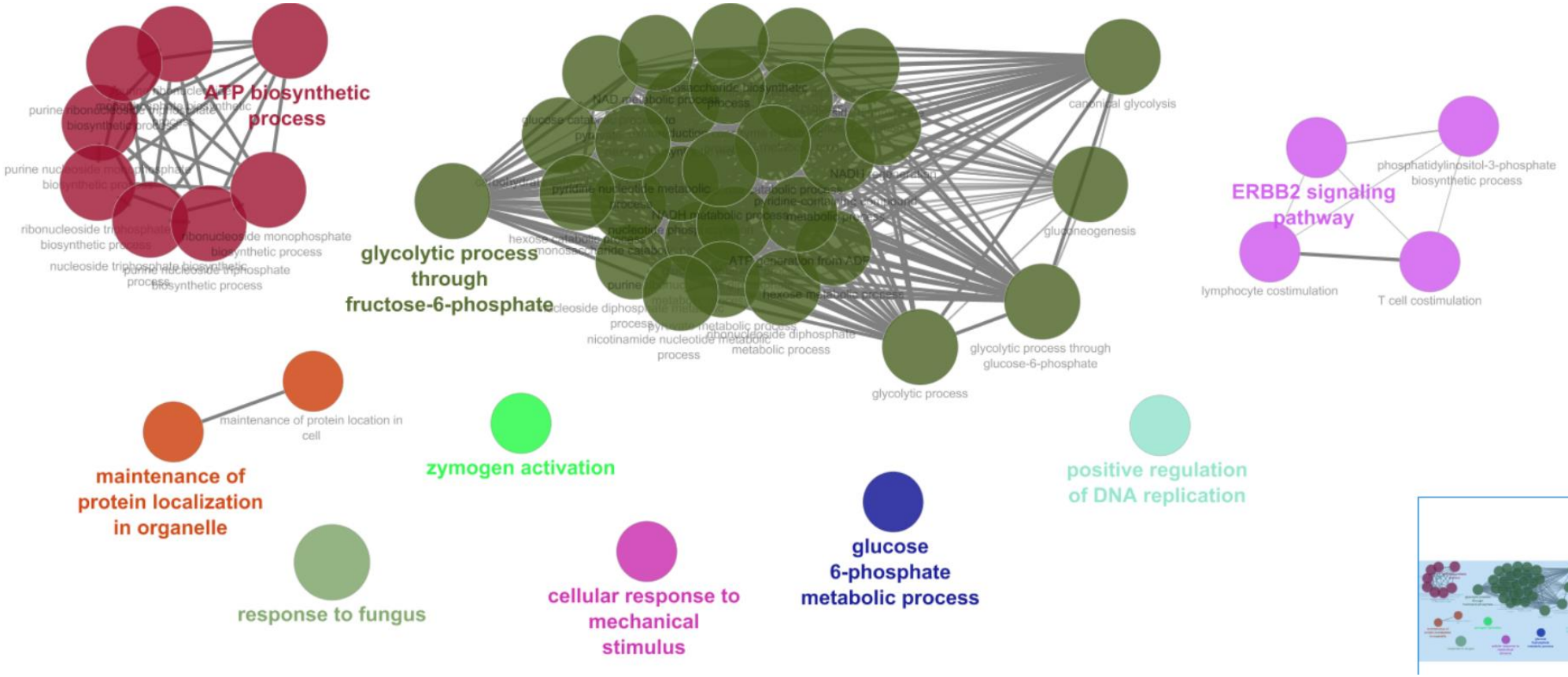
Pathway enrichment analysis software: Cytoscape / ClueGO



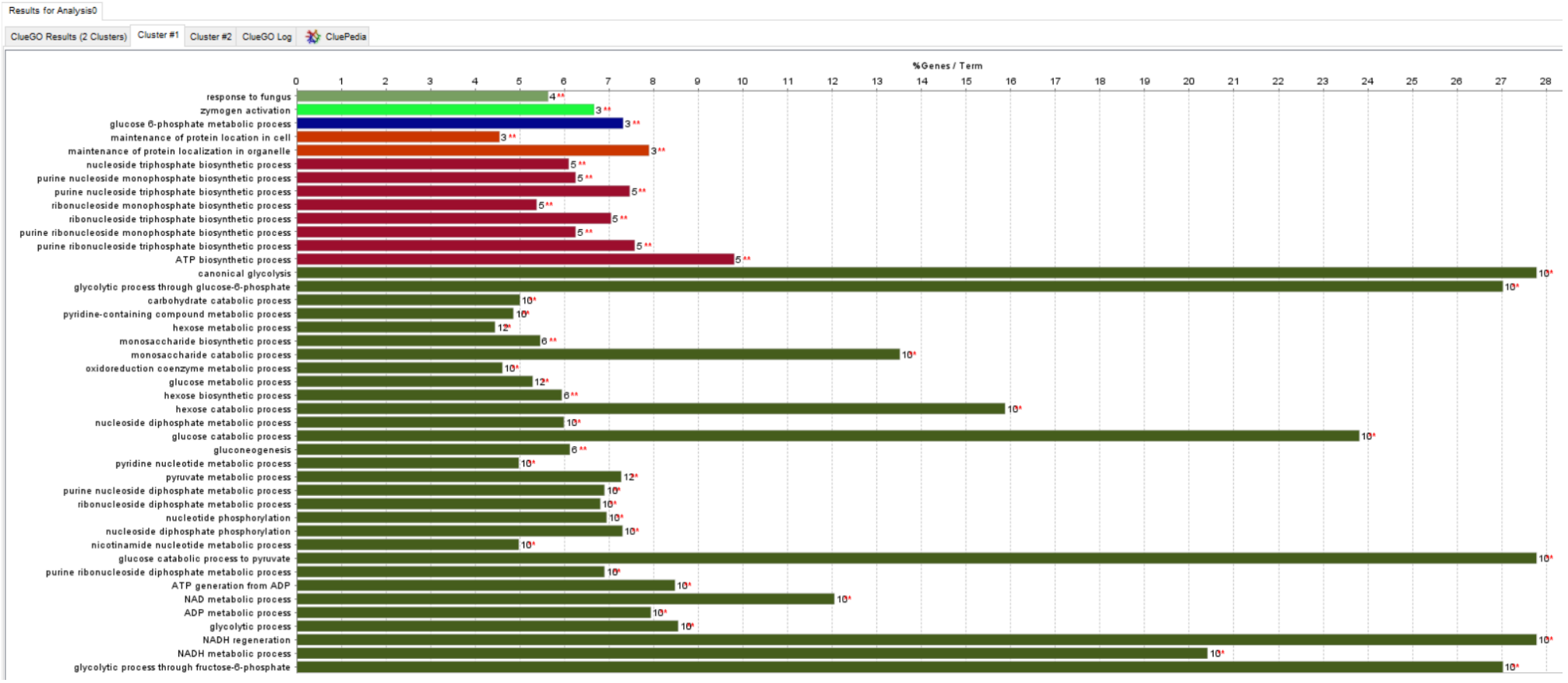
Pathway enrichment analysis software: Cytoscape / ClueGO



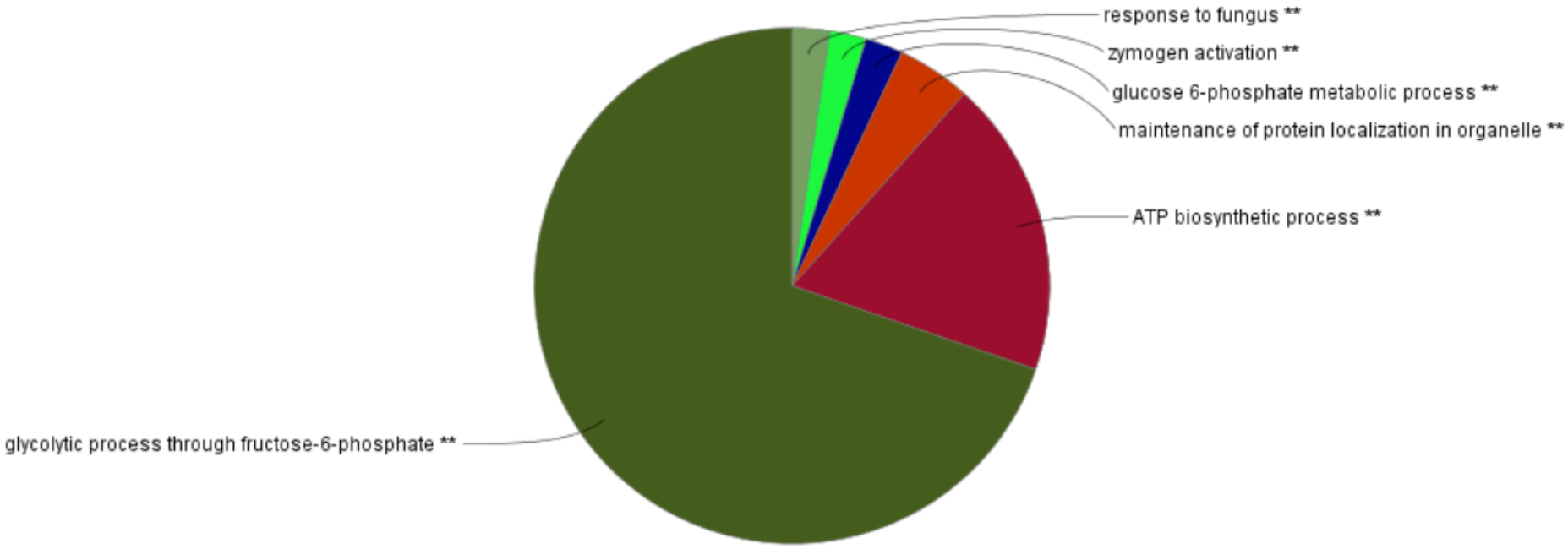
Pathway enrichment analysis software: Cytoscape / ClueGO



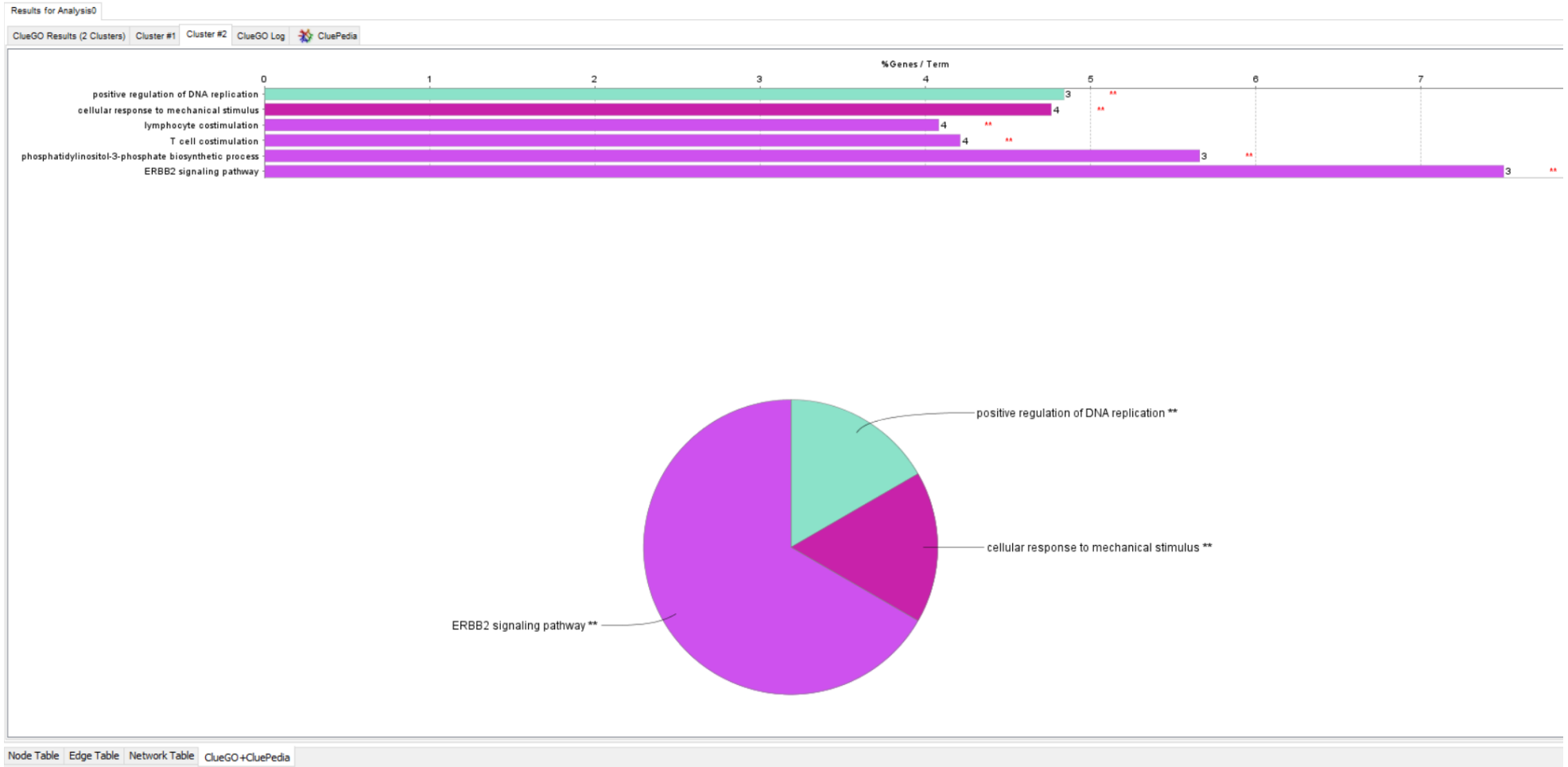
Pathway enrichment analysis software: Cytoscape / ClueGO



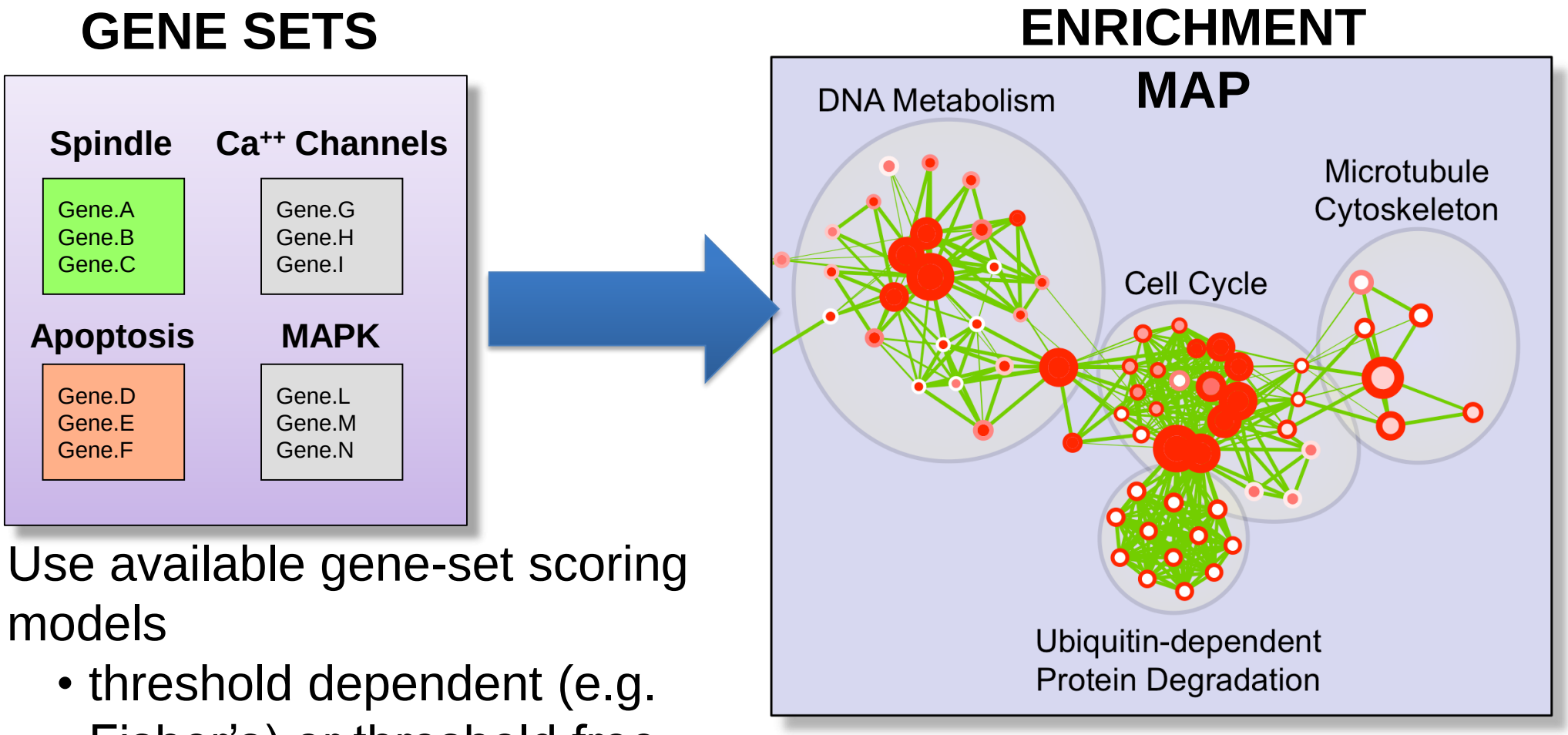
Pathway enrichment analysis software: Cytoscape / ClueGO



Pathway enrichment analysis software: Cytoscape / ClueGO

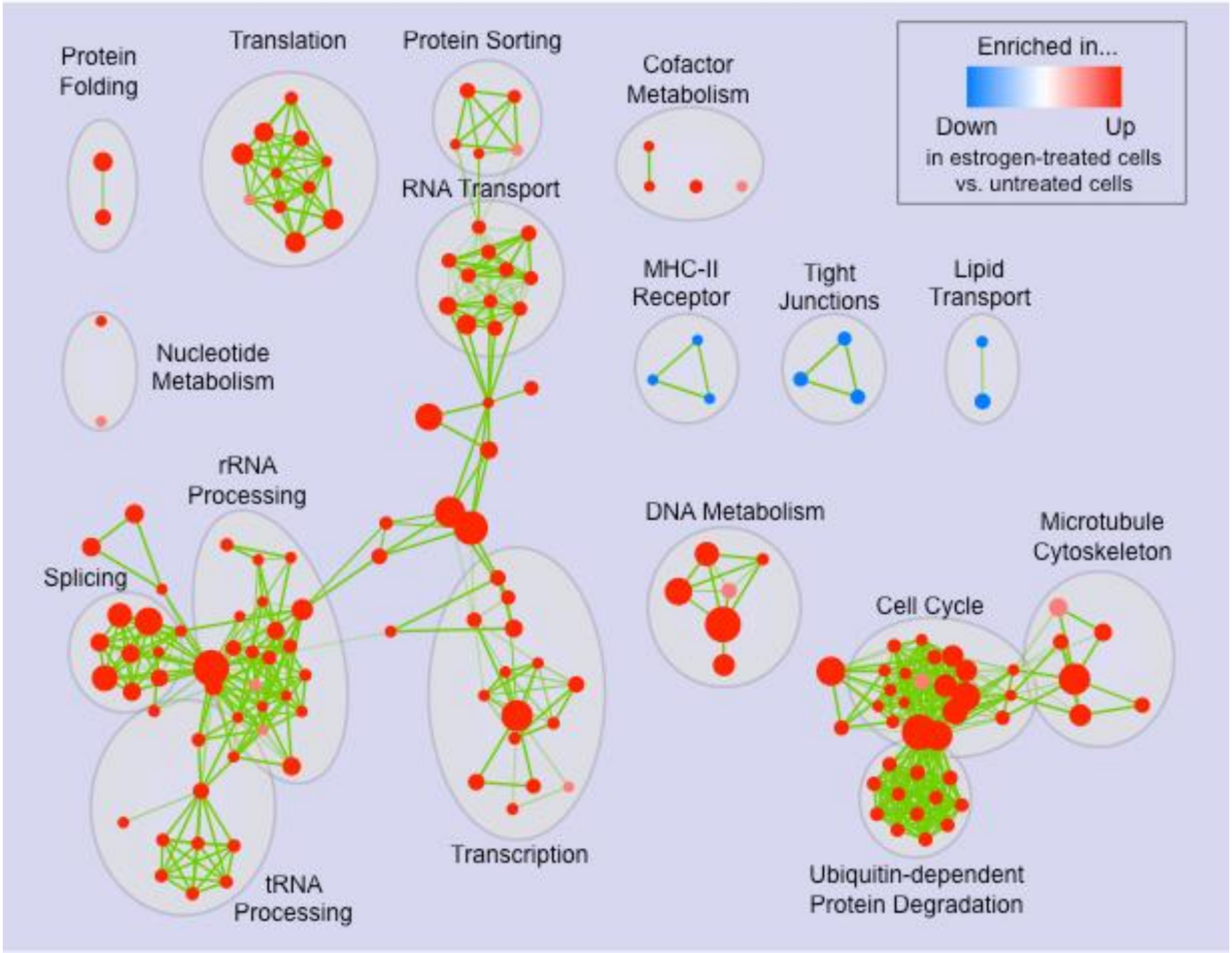


Enrichment Map



- Use available gene-set scoring models
  - threshold dependent (e.g. Fisher's) or threshold free (e.g. GSEA)
- Use the network framework to organize gene-sets exploiting their inter-dependencies

http://baderlab.org/Software/EnrichmentMap/


Protein Folding
Translation
Protein Sorting
Cofactor Metabolism
Enriched in...
Down
Up
in estrogen-treated cells vs. untreated cells
RNA Transport
MHC-II Receptor
Tight Junctions
Lipid Transport
Nucleotide Metabolism
rRNA Processing
Splicing
DNA Metabolism
Microtubule Cytoskeleton
Cell Cycle
Transcription
tRNA Processing
Ubiquitin-dependent Protein Degradation

Pathway enrichment analysis software: R / Bioconductor

- piano
- EGSEA
- ToPASeq
- GSVA
- GAGE
- PGSEA
- SeqGSEA
- sigPathway
- And many more

Final remarks:

- You can always find standalone and web-based applications for pathway analysis, but many tools exist either as scripts or as libraries that you must run.
- Therefore, it is good to learn how to program.
- Currently, the two most popular programming languages in bioinformatics are **R** and **python**. R has a suite of software for bioinformatics called "**Bioconductor**", while python has "**bioconda**".
- Learn R!

What have we learned today?

What is pathway/gene-set analysis
How to perform gene set analysis
Two types of gene set analysis (ORA and FCS)
What is multiple test correction
How to use software for gene set analysis (ORA and FCS)


MORA LAB
FUNCTIONAL BIOINFORMATICS