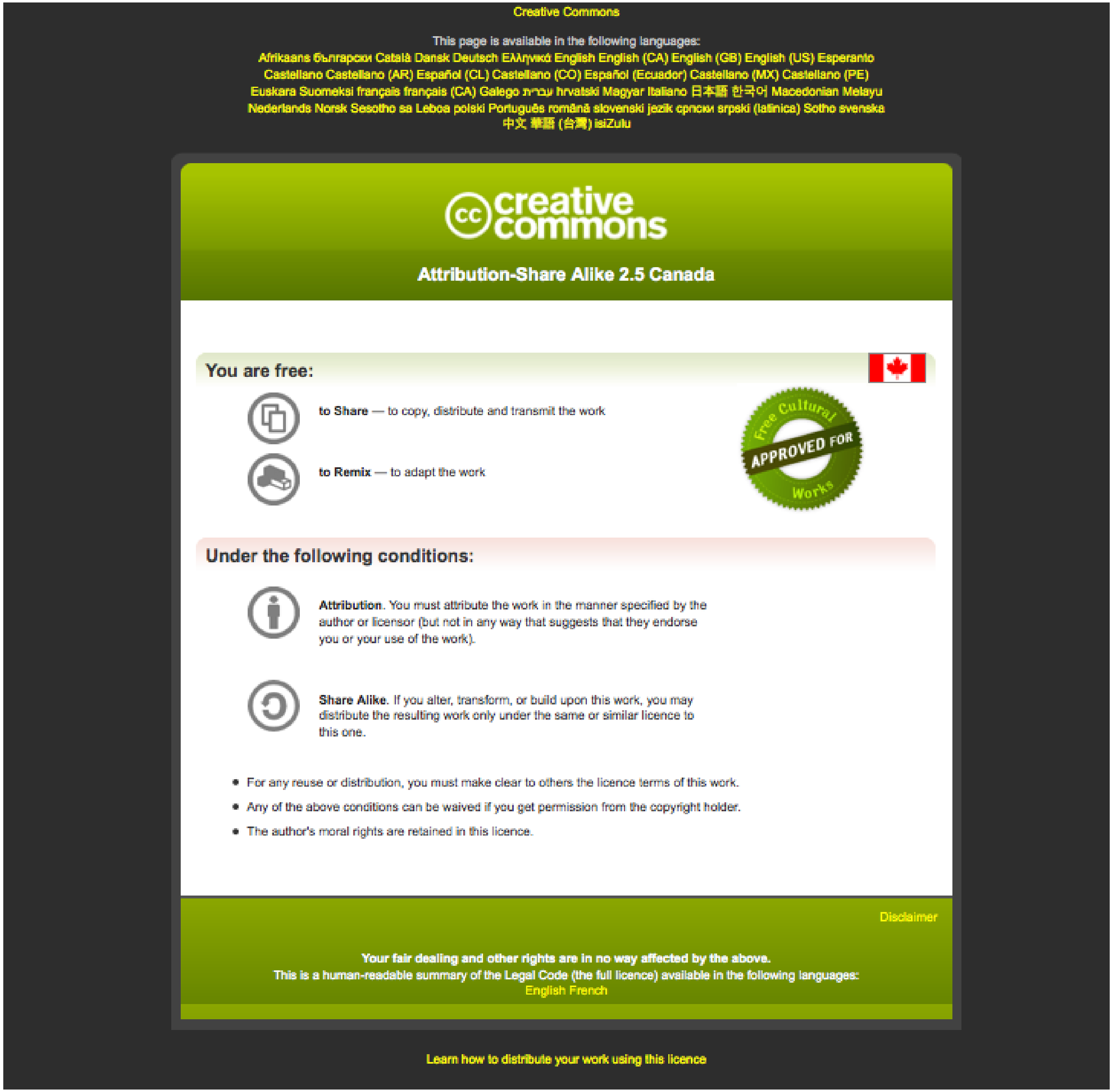
廣州醫科大學
GUANGZHOU MEDICAL UNIVERSITY

Gene Set Analysis –Methods and Tools

Antonio Mora, Ph.D.
(Wechat: antoniocmora)
www.moralab.science

21.09.2020

MORA LAB
FUNCTIONAL BIOINFORMATICS



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

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

Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

Contents

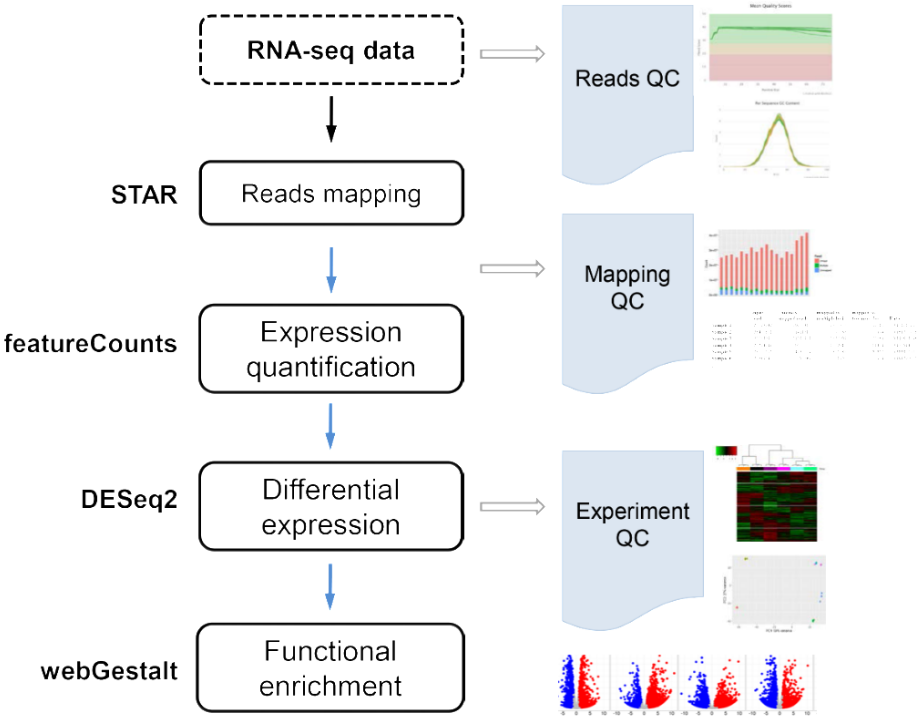
2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

GSA is interpretation of results in terms of gene sets

You may have heard about:

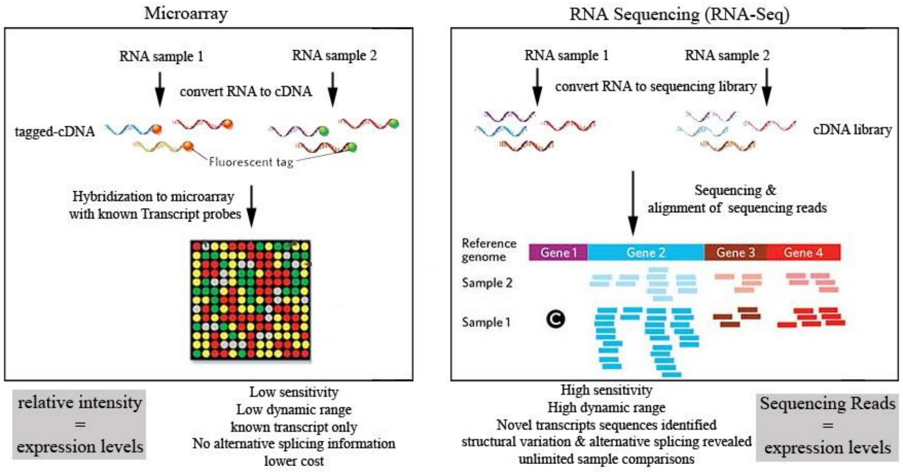
- Pathway (enrichment) Analysis
- Gene Set (enrichment) Analysis
- Functional Enrichment Analysis
- Ontology Analysis
- Knowledge-driven pathway analysis
- And other names…

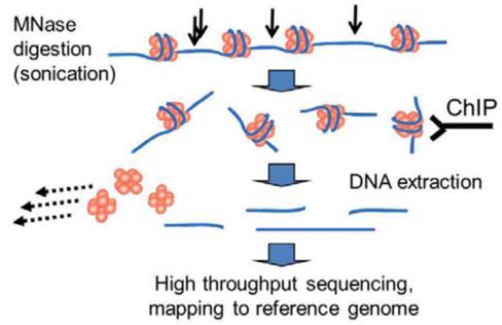
It is all the same. We are at the end of a research project and we want to find the meaning of the group of biological molecules that we obtained as a result. What is interesting about them? How are they related to each other?



http://bioinfo.vanderbilt.edu/vangard/services-rnaseq.html

GSA is interpretation of results in terms of gene sets





ChIP-seq workflow and data analysis.

TYPES OF EXPERIMENTS:

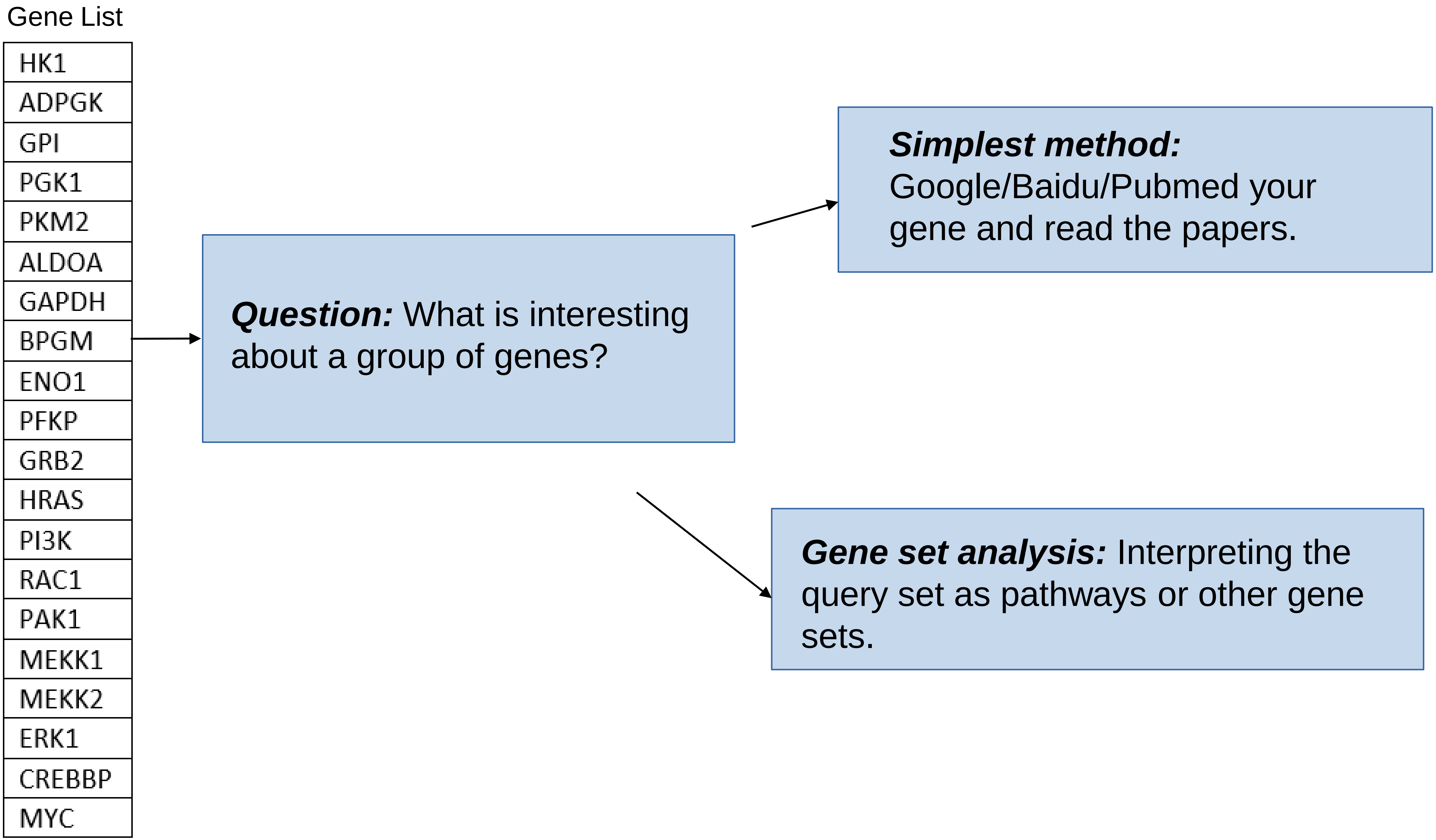
- Molecular profiling (mRNA, protein)
- Interactions (TF binding sites, miRNA targets)
- Association studies (SNPs, CNVs)

RESULTS →

Gene List
HK1
ADPGK
GPI
PGK1
PKM2
ALDOA
GAPDH
BPGM
ENO1
PFKP
GRB2
HRAS
PI3K
RAC1
PAK1
MEKK1
MEKK2
ERK1
CREBBP
MYC

https://www.otogenetics.com/wp-content/uploads/2017/12/RNA-Seq-VS-Microarray.jpg
https://www.researchgate.net/figure/ChIP-seq-workflow-and-data-analysis_fig1_321662815

GSA is interpretation of results in terms of gene sets



Gene List

HK1
ADPGK
GPI
PGK1
PKM2
ALDOA
GAPDH
BPGM
ENO1
PFKP
GRB2
HRAS
PI3K
RAC1
PAK1
MEKK1
MEKK2
ERK1
CREBBP
MYC

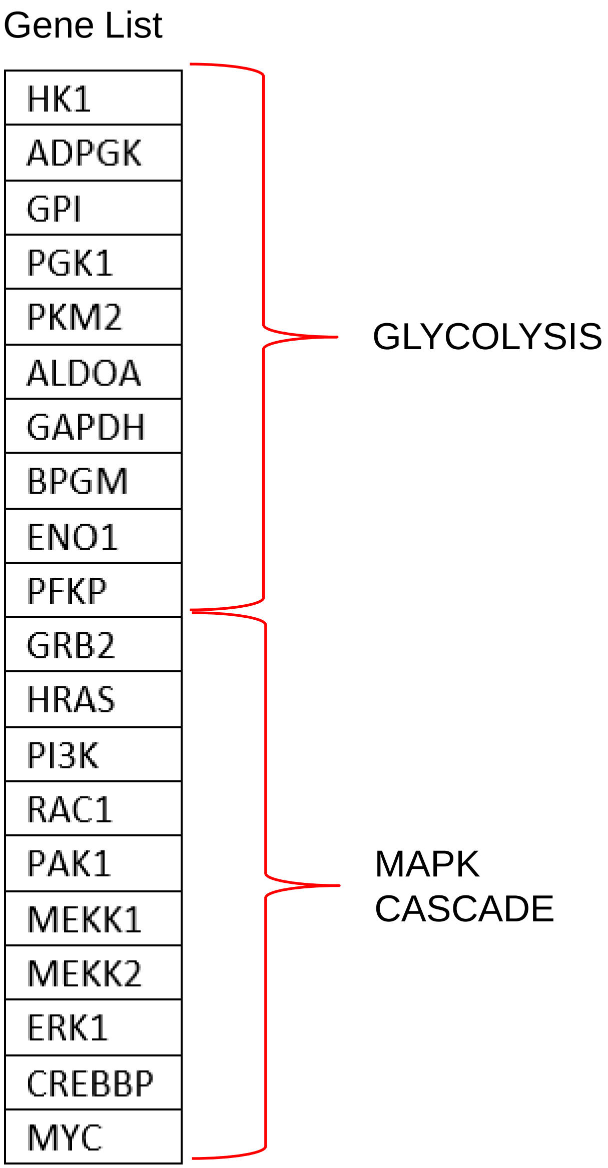
Question: What is interesting about a group of genes?

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

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

GSA is interpretation of results in terms of gene sets

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



Gene List

HK1
ADPGK
GPI
PGK1
PKM2
ALDOA
GAPDH
BPGM
ENO1
PFKP
GRB2
HRAS
PI3K
RAC1
PAK1
MEKK1
MEKK2
ERK1
CREBBP
MYC

GLYCOLYSIS

MAPK CASCADE

Why Gene Set Analysis?

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

“Gene Set Analysis” Elements:

A query set: A group of genes that were the result of some experiment
Example of query set: Differentially expressed genes (up-regulated, down-regulated, or the entire list).

→

HK1
ADPGK
GPI
PGK1
PKM2
ALDOA
GAPDH
BPGM
ENO1
PFKP
GRB2
HRAS
PI3K
RAC1
PAK1
MEKK1
MEKK2
ERK1
CREBBP
MYC

Reference Databases:
Pathway / Ontology / Gene set Databases.

KEGG Kyoto Encyclopedia of Genes and Genomes

REACTOME A CURATED PATHWAY DATABASE

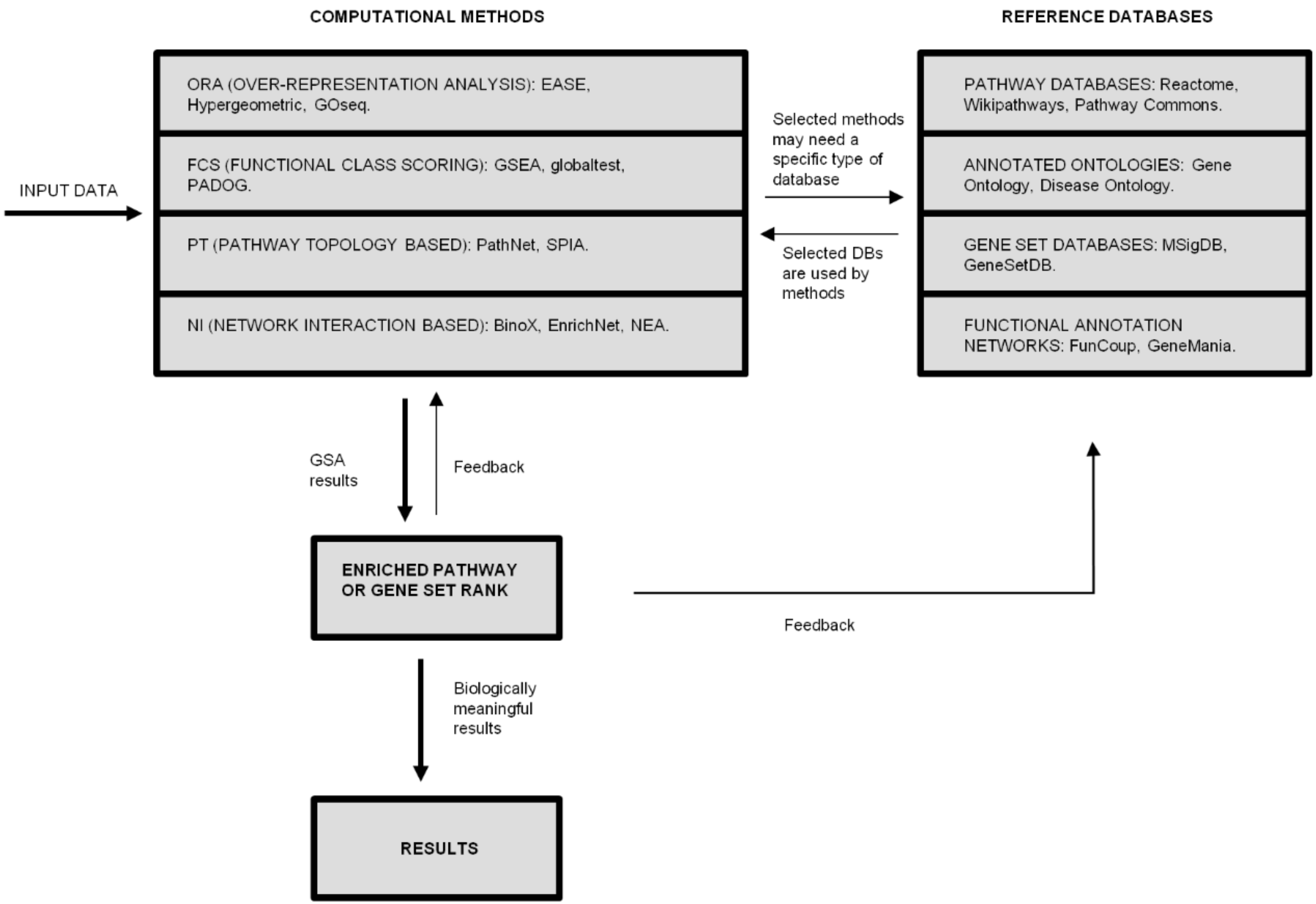
BETA WIKIPATHWAYS Pathways for the People

Statistical Method

→

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

Gene Set Analysis Workflow



COMPUTATIONAL METHODS

ORA (OVER-REPRESENTATION ANALYSIS): EASE, Hypergeometric, GOseq.

FCS (FUNCTIONAL CLASS SCORING): GSEA, globaltest, PADOG.

PT (PATHWAY TOPOLOGY BASED): PathNet, SPIA.

NI (NETWORK INTERACTION BASED): BinoX, EnrichNet, NEA.

REFERENCE DATABASES

PATHWAY DATABASES: Reactome, Wikipathways, Pathway Commons.

ANNOTATED ONTOLOGIES: Gene Ontology, Disease Ontology.

GENE SET DATABASES: MSigDB, GeneSetDB.

FUNCTIONAL ANNOTATION NETWORKS: FunCoup, GeneMania.

INPUT DATA

Selected methods may need a specific type of database

Selected DBs are used by methods

GSA results

Feedback

ENRICHED PATHWAY OR GENE SET RANK

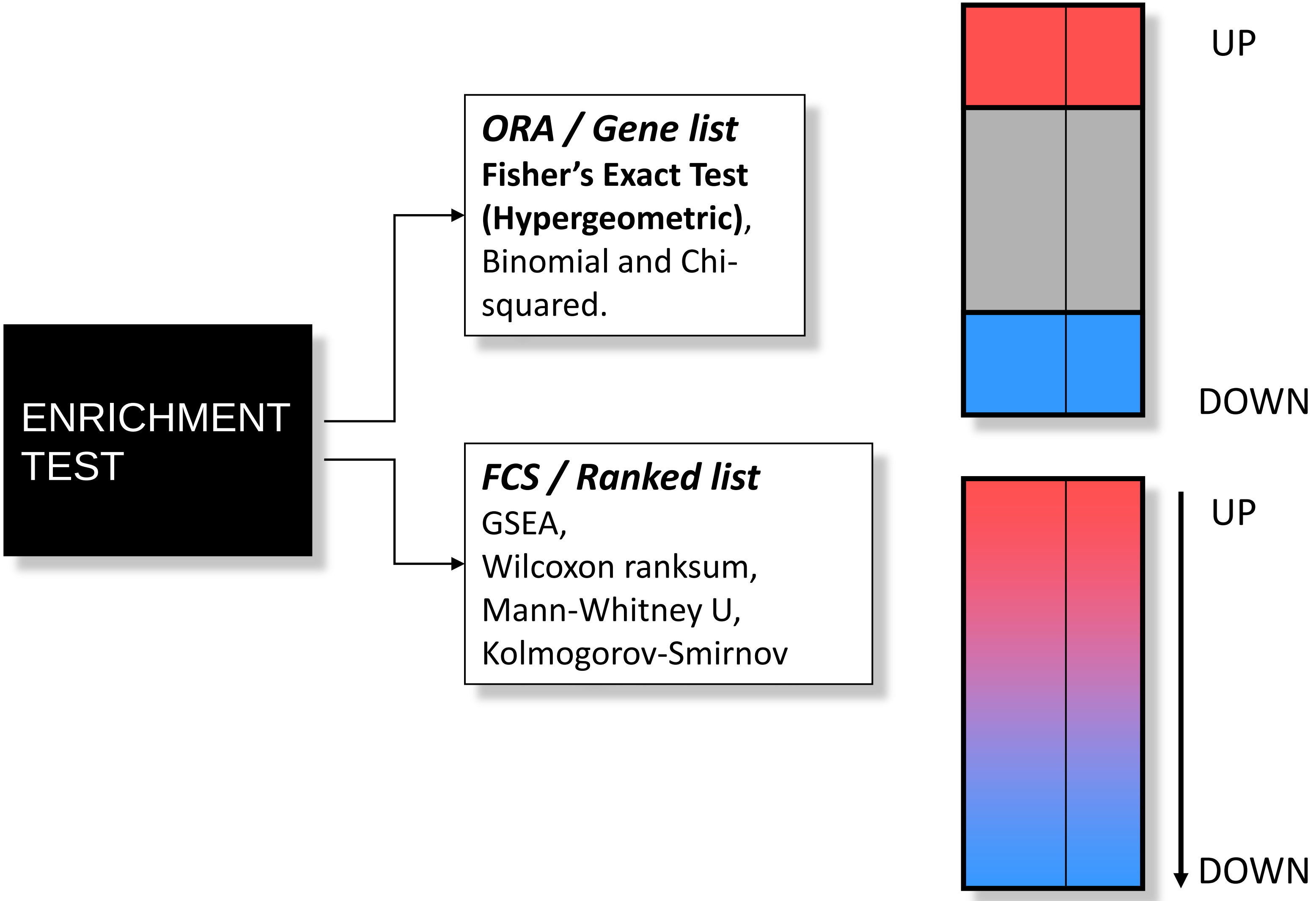
Feedback

Biologically meaningful results

RESULTS

https://doi.org/10.1093/bib/bbz090

Statistical Tests



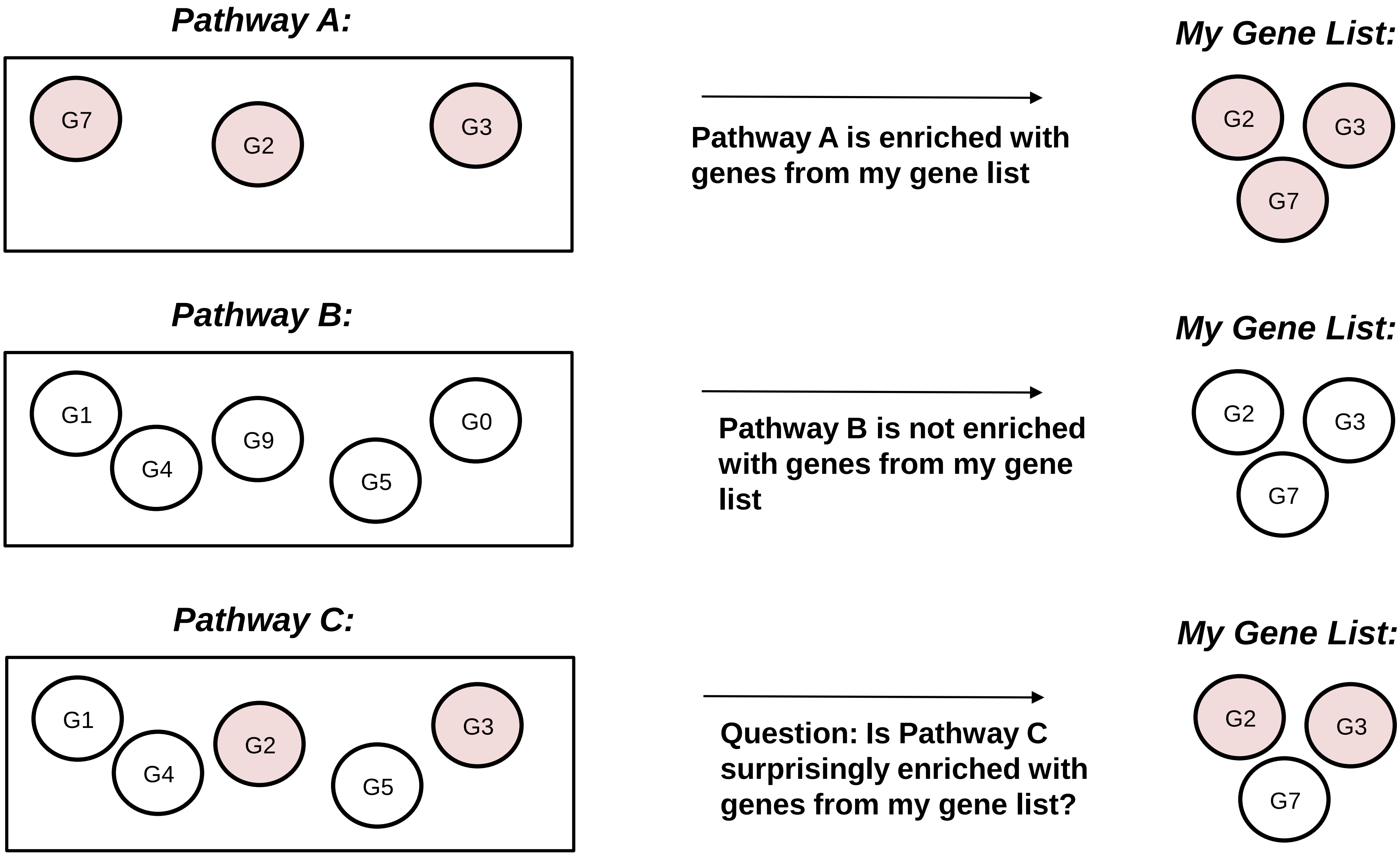
ENRICHMENT TEST

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

The ORA approach (For a gene list, e.g. genes with expression change > 2-fold)



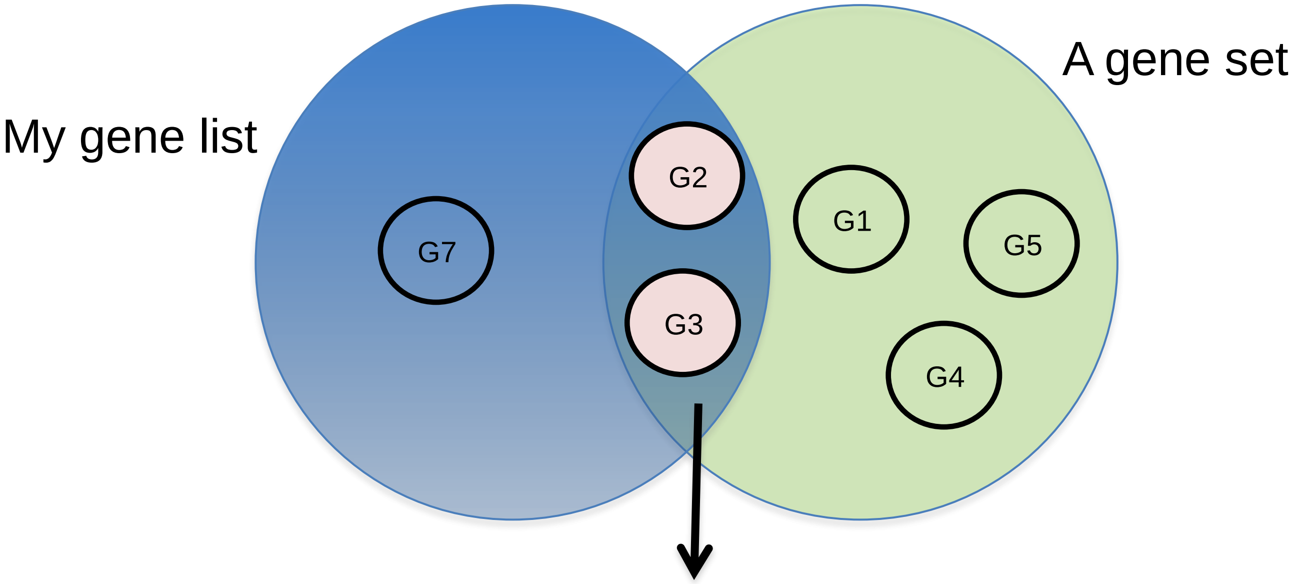
Pathway A is enriched with genes from my gene list

Pathway B is not enriched with genes from my gene list

Question: Is Pathway C surprisingly enriched with genes from my gene list?

The ORA approach (For a gene list, e.g. genes with expression change > 2-fold)

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.



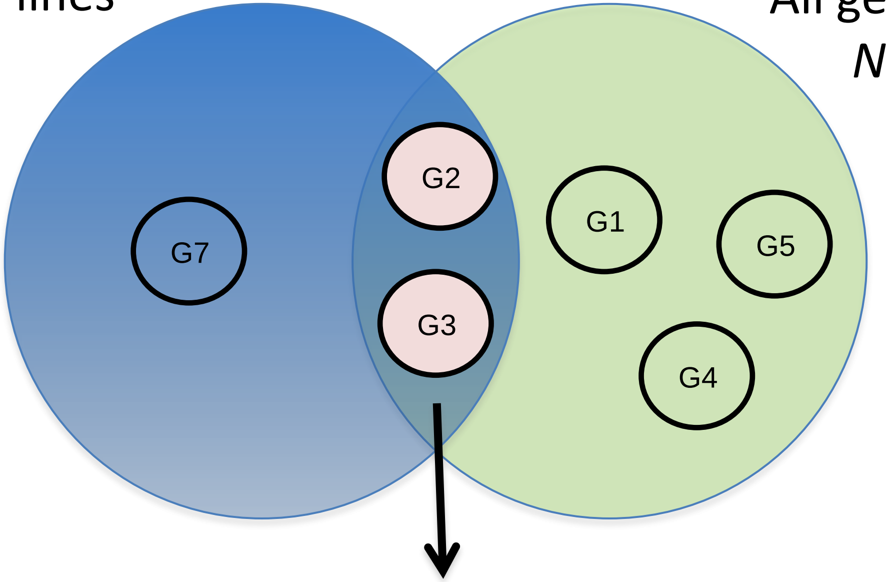
Are the genes in the intersection too many? What do we mean when we say “too many”? 5 out of 10? 7 out of 10? (We must use Statistics and compare to how many we can find by chance alone!)

The ORA approach (For a gene list, e.g. genes with expression change > 2-fold)

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

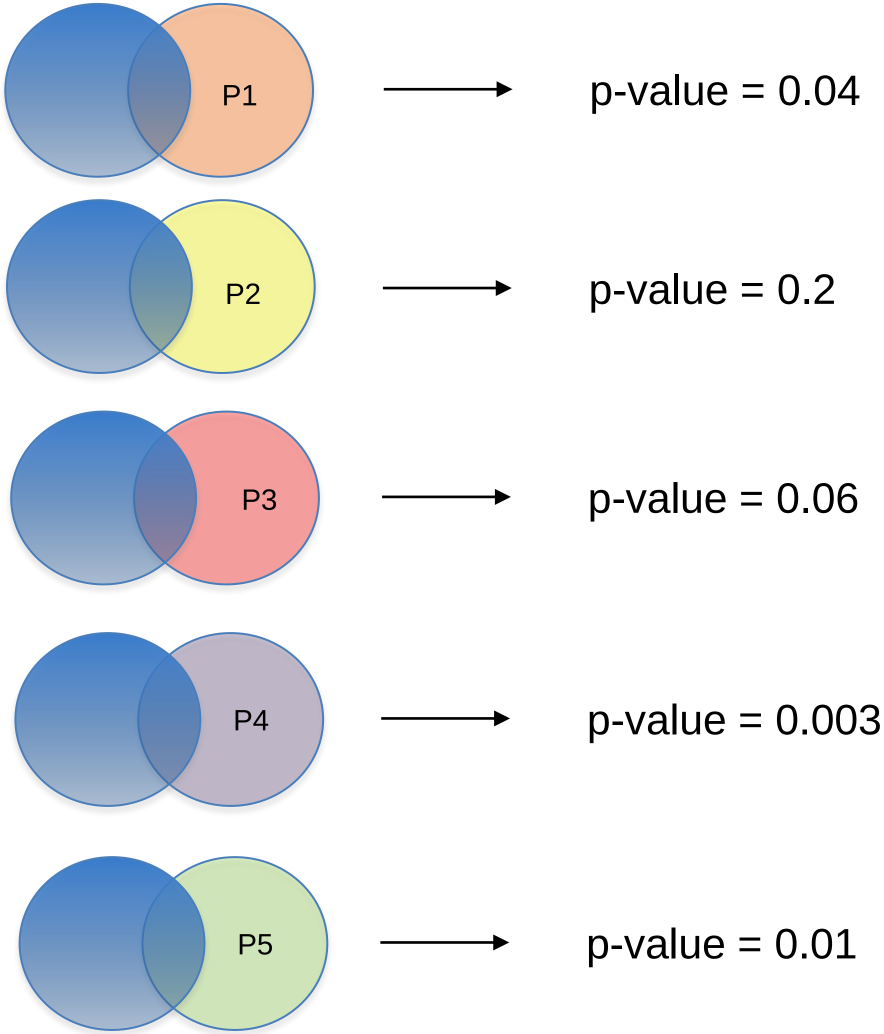
Pathway information:
All genes in the pathway called *Neurotransmitter signaling*



Statistical test: Are there more genes in the intersection than expected by chance alone? (p-value < 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

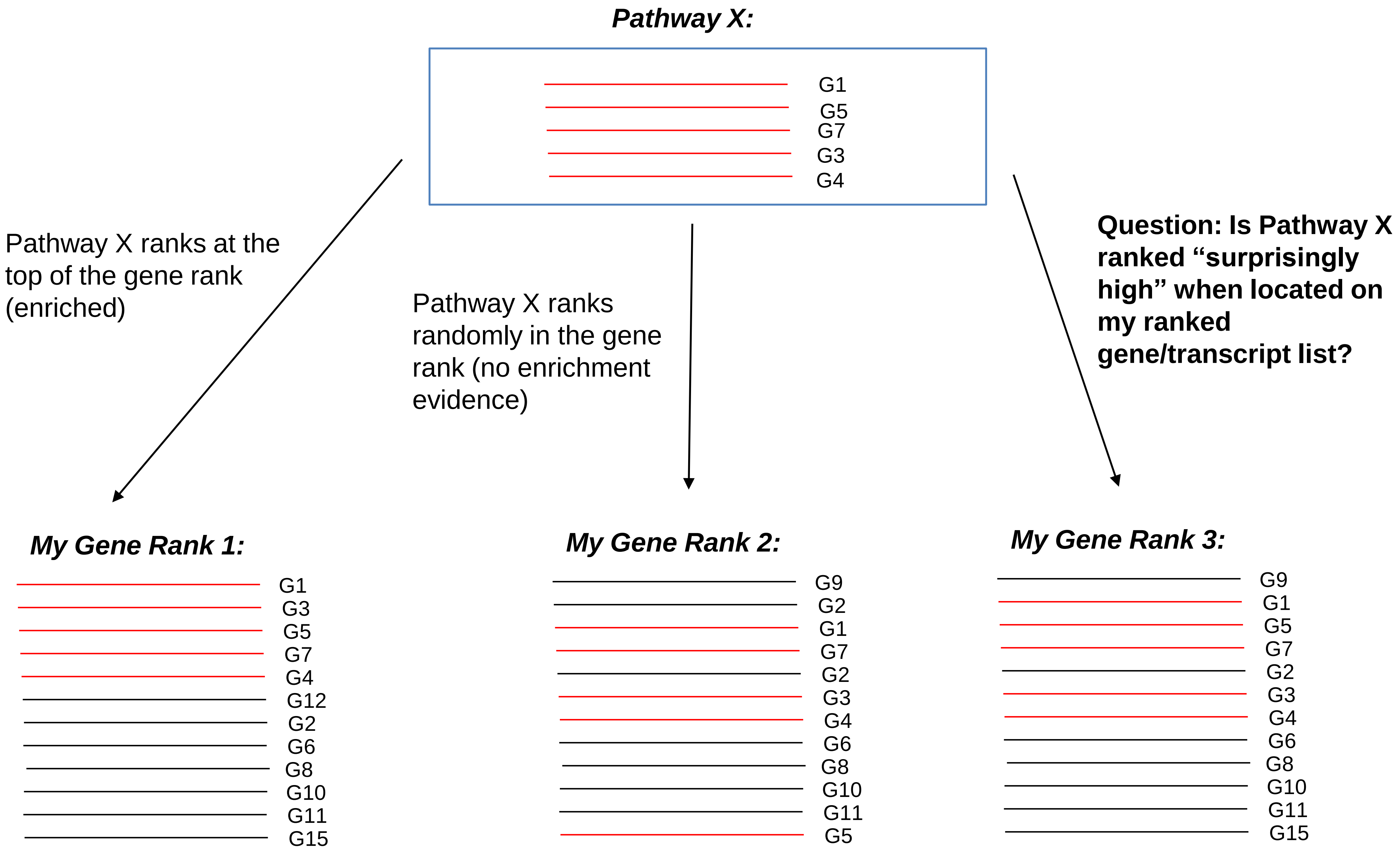
Gene Set	p-value
P4	0.003
P5	0.01
P1	0.04
P3	0.06
P2	0.2

Significant

Cutoff

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

The FCS approach (Gene rank, e.g. entire list, ordered by differential expression)



Pathway X:

G1
G5
G7
G3
G4

Pathway X ranks at the top of the gene rank (enriched)

Pathway X ranks randomly in the gene rank (no enrichment evidence)

Question: Is Pathway X ranked “surprisingly high” when located on my ranked gene/transcript list?

My Gene Rank 1:

G1
G3
G5
G7
G4
G12
G2
G6
G8
G10
G11
G15

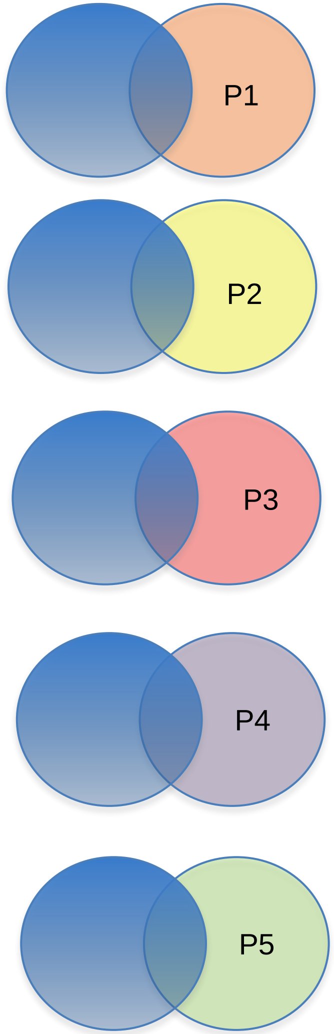
My Gene Rank 2:

G9
G2
G1
G7
G2
G3
G4
G6
G8
G10
G11
G5

My Gene Rank 3:

G9
G1
G5
G7
G2
G3
G4
G6
G8
G10
G11
G15

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

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?

Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

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

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

HK1
ADPGK
GPI
PGK1
PKM2
ALDOA
GAPDH
BPGM
ENO1
PFKP
GRB2
HRAS
PI3K
RAC1
PAK1
MEKK1
MEKK2
ERK1
CREBBP
MYC

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

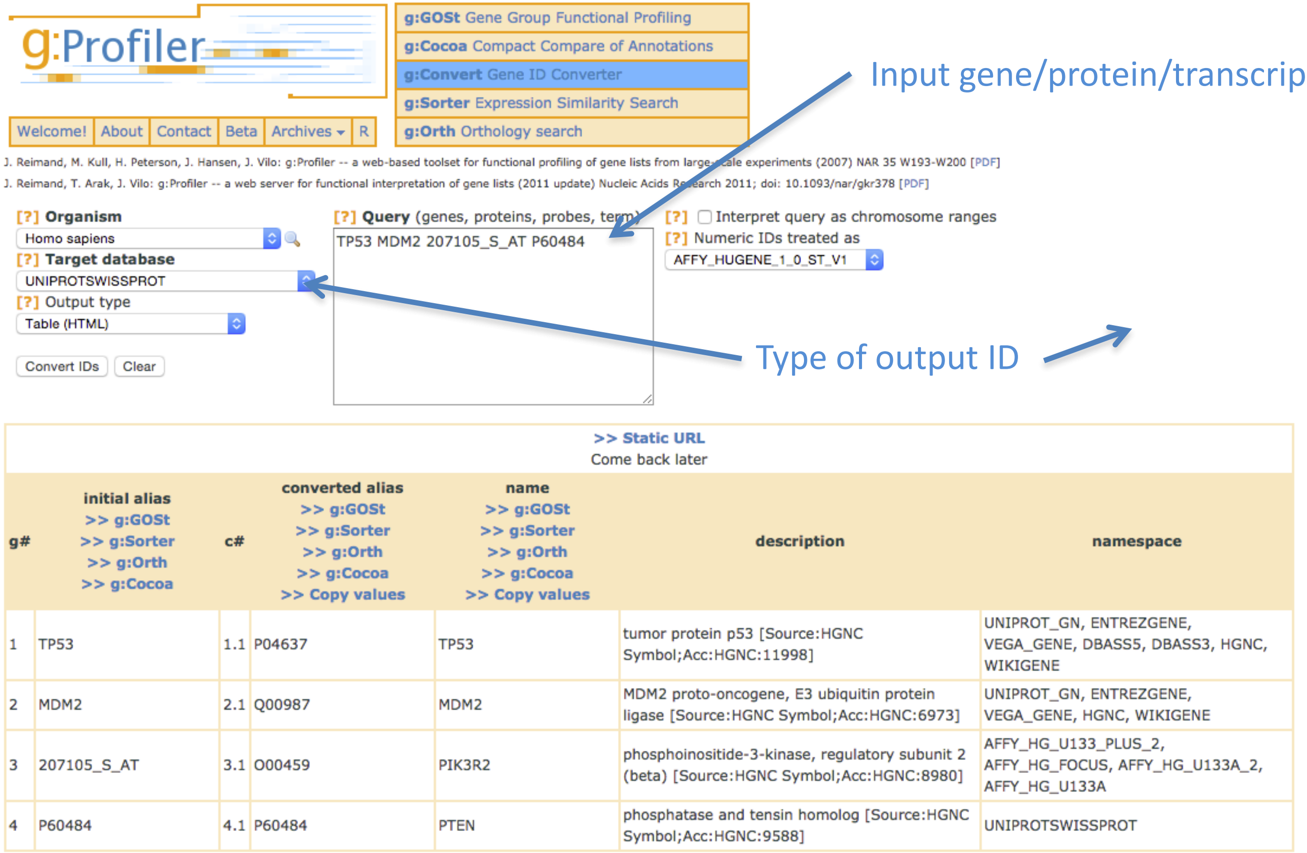
Experimental Platform
Affymetrix 208368_3p_s_at
Agilent A_23_P99452
CodeLink GE60169
Illumina GI_4502450-S

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

g:Profiler
Welcome! | About | Contact | Beta | Archives ▾ | R
g:GOSt Gene Group Functional Profiling
g:Cocoa Compact Compare of Annotations
g:Convert Gene ID Converter
g:Sorter Expression Similarity Search
g:Orth Orthology search

J. Reimand, M. Kull, H. Peterson, J. Hansen, J. Vilo: g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments (2007) NAR 35 W193-W200 [PDF]

J. Reimand, T. Arak, J. Vilo: g:Profiler -- a web server for functional interpretation of gene lists (2011 update) Nucleic Acids Research 2011; doi: 10.1093/nar/gkr378 [PDF]

[?] Organism: Homo sapiens

[?] Target database: UNIPROTSWISSPROT

[?] Output type: Table (HTML)

Convert IDs | Clear

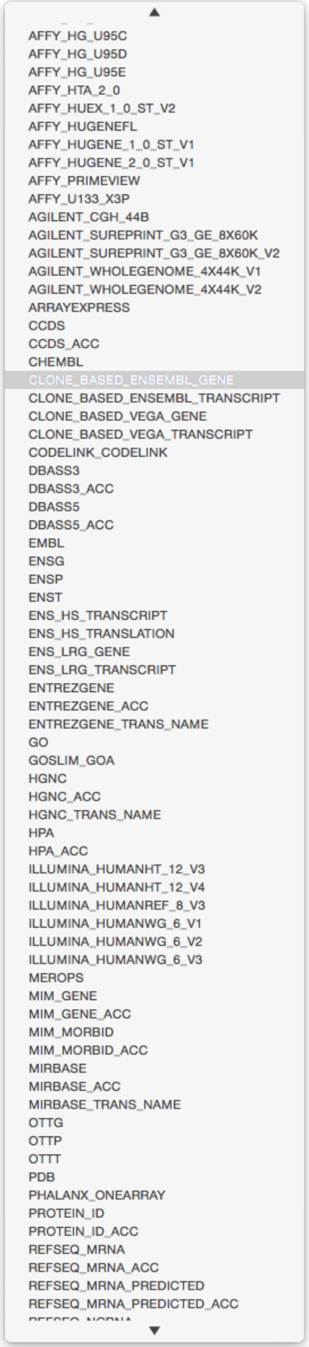
[?] Query (genes, proteins, probes, term): TP53 MDM2 207105_S_AT P60484

[?] ☐ Interpret query as chromosome ranges

[?] Numeric IDs treated as: AFFY_HUGENE_1_0_ST_V1

>> 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



AFFY_HG_U95C
AFFY_HG_U95D
AFFY_HG_U95E
AFFY_HTA_2_0
AFFY_HUEX_1_0_ST_V2
AFFY_HUGENEFL
AFFY_HUGENE_1_0_ST_V1
AFFY_HUGENE_2_0_ST_V1
AFFY_PRIMEVIEW
AFFY_U133_X3P
AGILENT_CGH_44B
AGILENT_SUREPRINT_G3_GE_8X60K
AGILENT_SUREPRINT_G3_GE_8X60K_V2
AGILENT_WHOLEGENOME_4X44K_V1
AGILENT_WHOLEGENOME_4X44K_V2
ARRAYEXPRESS
CCDS
CCDS_ACC
CHEMBL
CLONE_BASED_ENSEMBL_GENE
CLONE_BASED_ENSEMBL_TRANSCRIPT
CLONE_BASED_VEGA_GENE
CLONE_BASED_VEGA_TRANSCRIPT
CODELINK_CODELINK
DBASS3
DBASS3_ACC
DBASS5
DBASS5_ACC
EMBL
ENSG
ENSP
ENST
ENS_HS_TRANSCRIPT
ENS_HS_TRANSLATION
ENS_LRG_GENE
ENS_LRG_TRANSCRIPT
ENTREZGENE
ENTREZGENE_ACC
ENTREZGENE_TRANS_NAME
GO
GOSLIM_GOA
HGNC
HGNC_ACC
HGNC_TRANS_NAME
HPA
HPA_ACC
ILLUMINA_HUMANHT_12_V3
ILLUMINA_HUMANHT_12_V4
ILLUMINA_HUMANREF_8_V3
ILLUMINA_HUMANWG_6_V1
ILLUMINA_HUMANWG_6_V2
ILLUMINA_HUMANWG_6_V3
MEROPS
MIM_GENE
MIM_GENE_ACC
MIM_MORBID
MIM_MORBID_ACC
MIRBASE
MIRBASE_ACC
MIRBASE_TRANS_NAME
OTTG
OTTP
OTTT
PDB
PHALANX_ONEARRAY
PROTEIN_ID
PROTEIN_ID_ACC
REFSEQ_MRNA
REFSEQ_MRNA_ACC
REFSEQ_MRNA_PREDICTED
REFSEQ_MRNA_PREDICTED_ACC

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

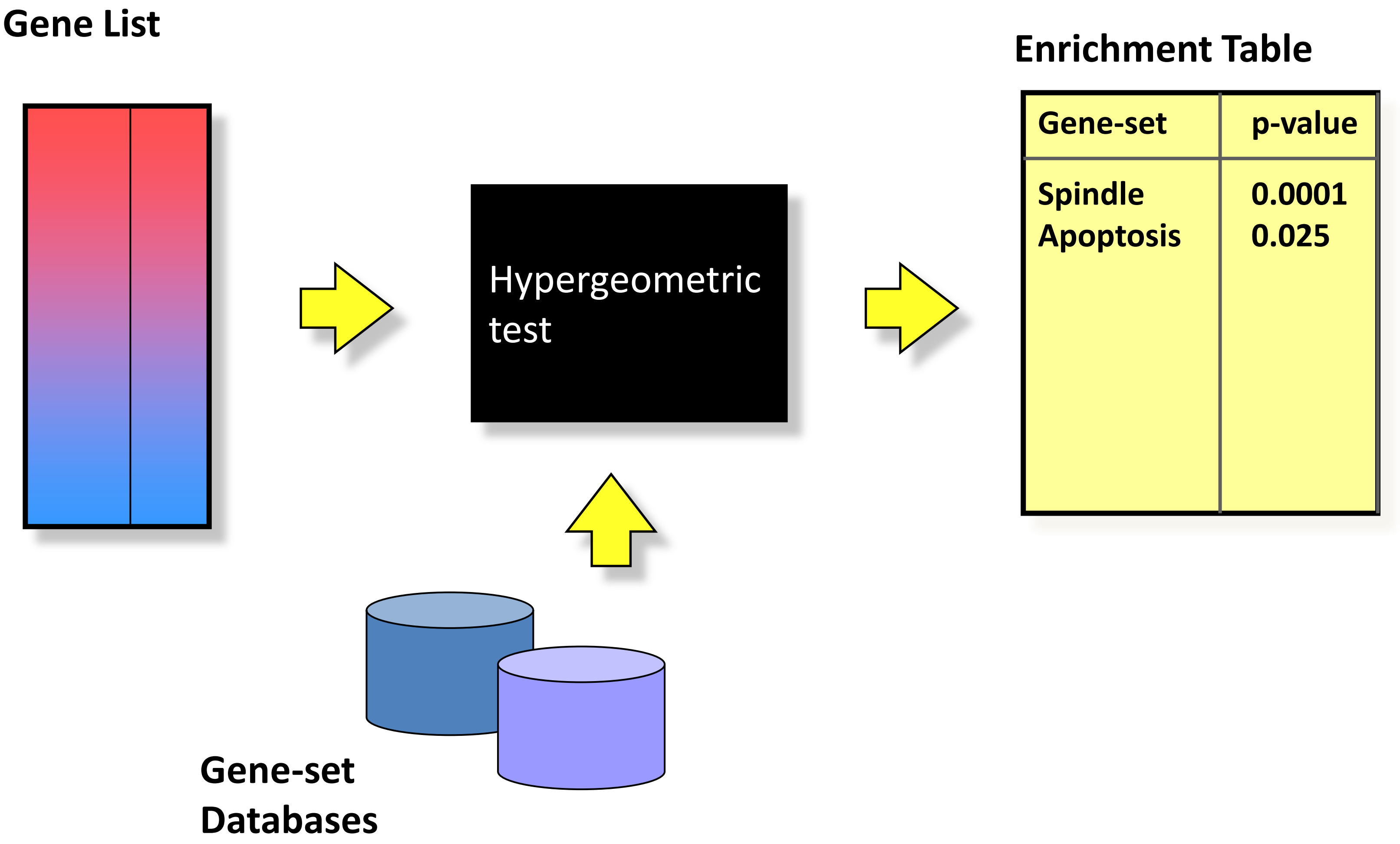
ID Challenges

- 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

Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software



Gene List

Hypergeometric test

Gene-set Databases

Enrichment Table

Gene-set	p-value
Spindle	0.0001
Apoptosis	0.025

Adapted from: Canadian Bioinformatics Workshop

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?

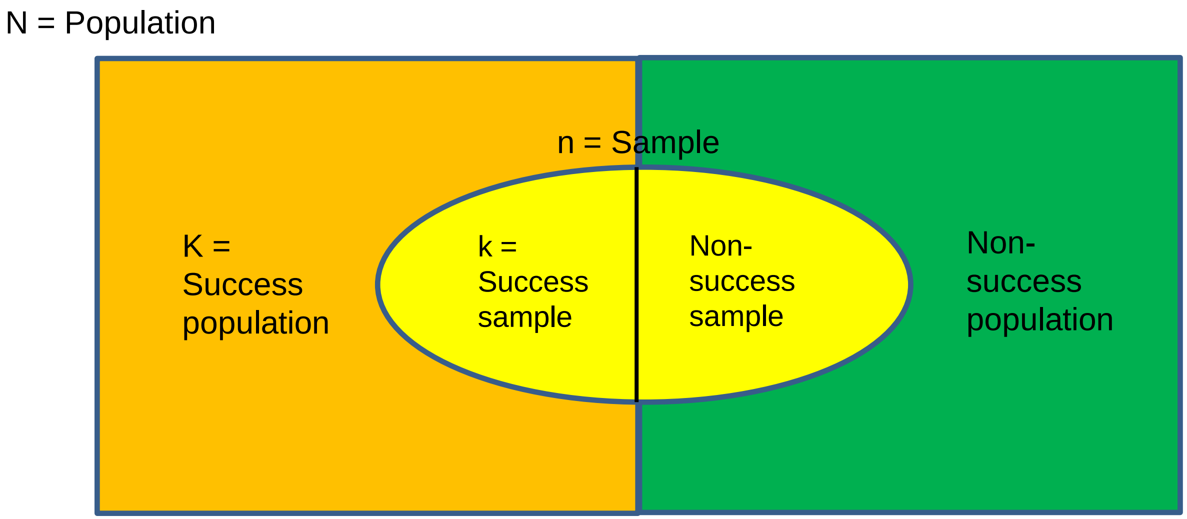
If it is very unlikely (low probability), we say that the sample genes are over-represented in that pathway.

PIC:

Low probability =
Difficult by chance
= Gene set may
represent gene list

High probability =
Easy by chance =
Gene set don’t
represent gene list

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?

Probability of success: P(X=k)

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

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

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

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

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

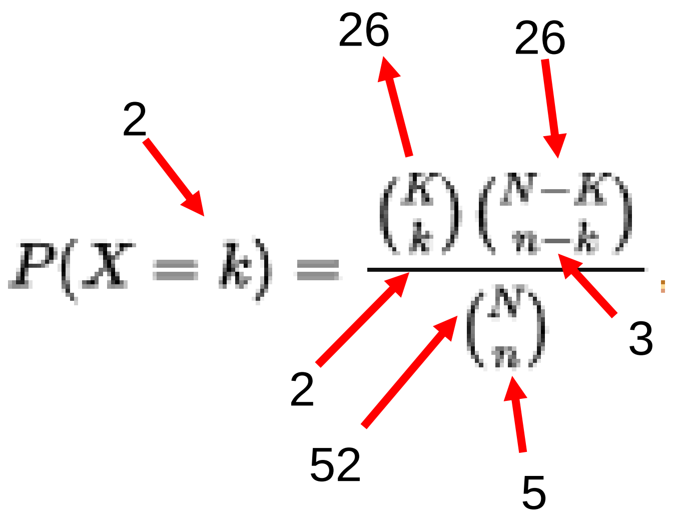
k = Sample success = 2

N – K = 26

n – k = 3

What is the probability of success?

Probability of success: P(X=k)



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

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

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

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.

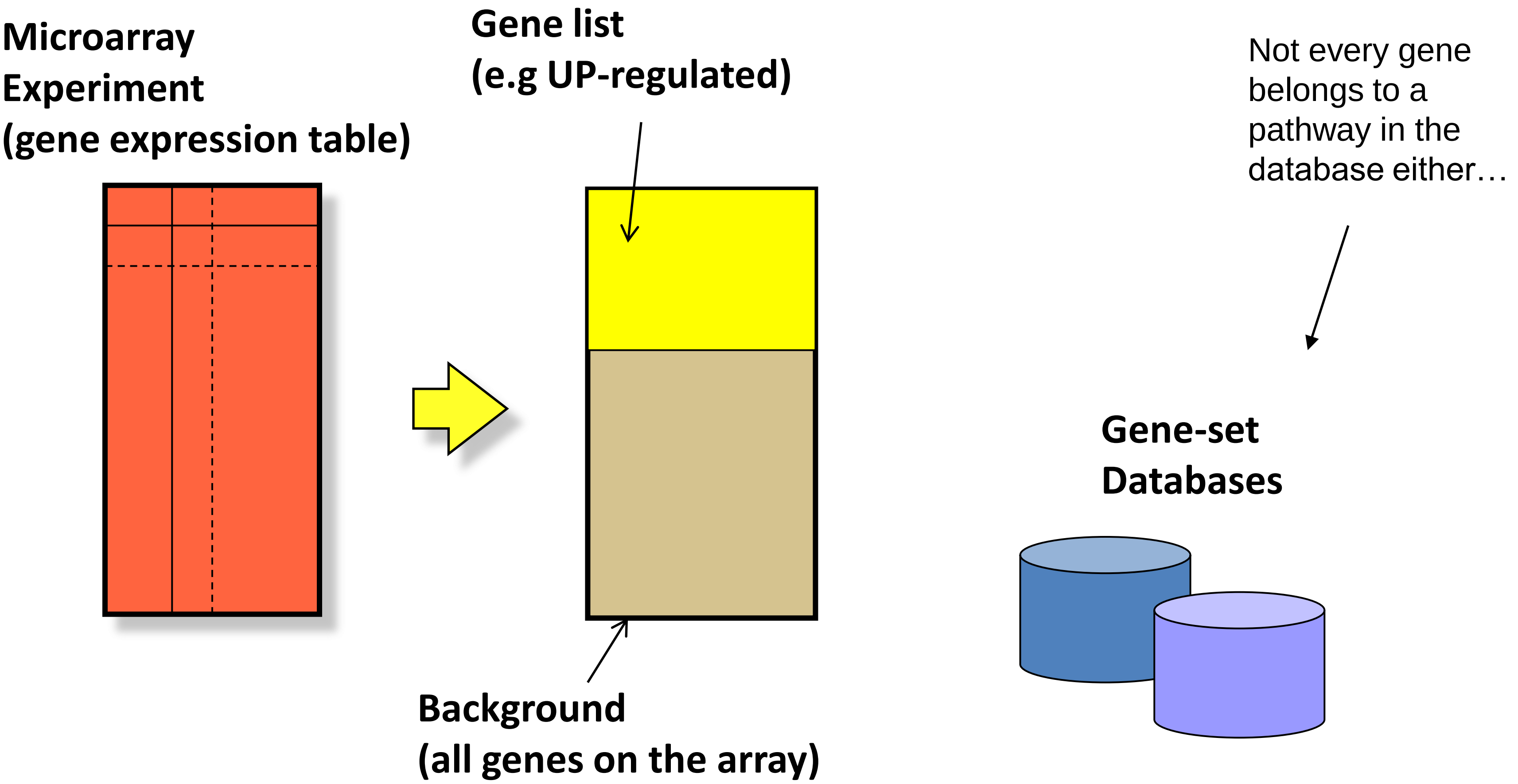
- ORA tools search for over-representation in a given database of pathways.

- 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.



Microarray Experiment
(gene expression table)

Gene list
(e.g UP-regulated)

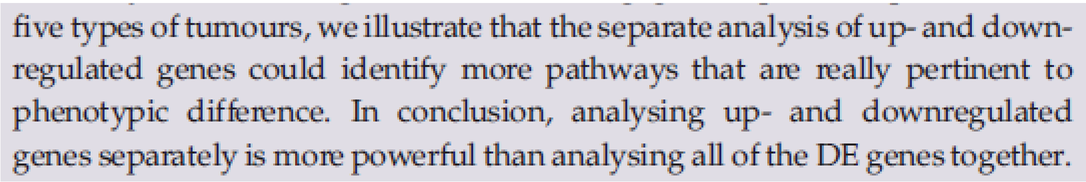
Not every gene belongs to a pathway in the database either…

Gene-set Databases

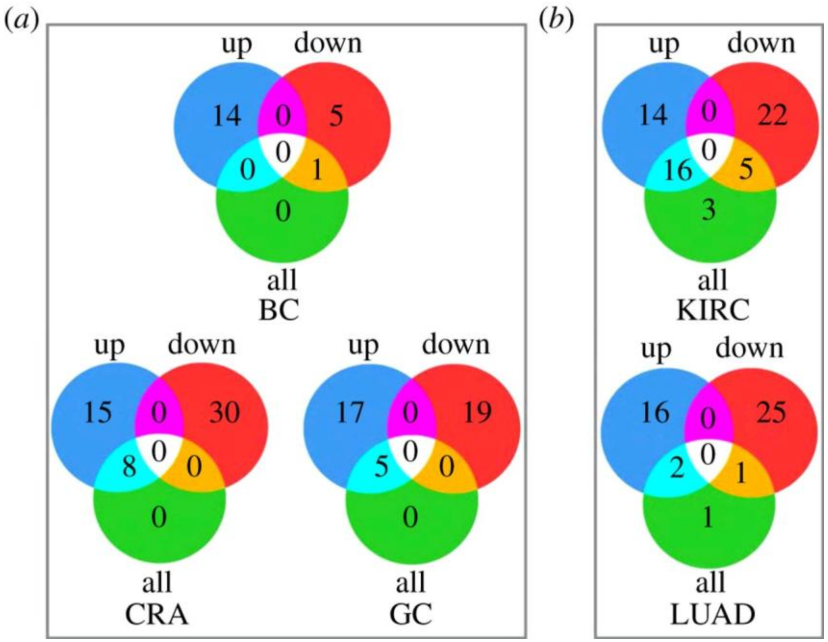
Background
(all genes on the array)

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 downregulated 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.



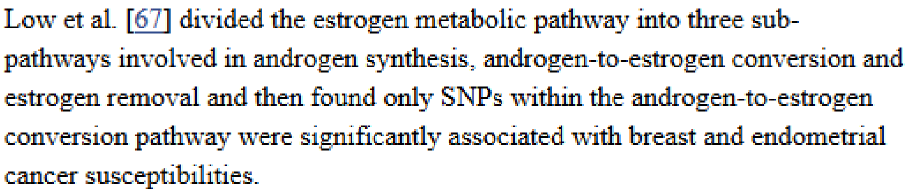
HONG G., Separate enrichment analysis of pathways for up- and downregulated genes, 2013

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.

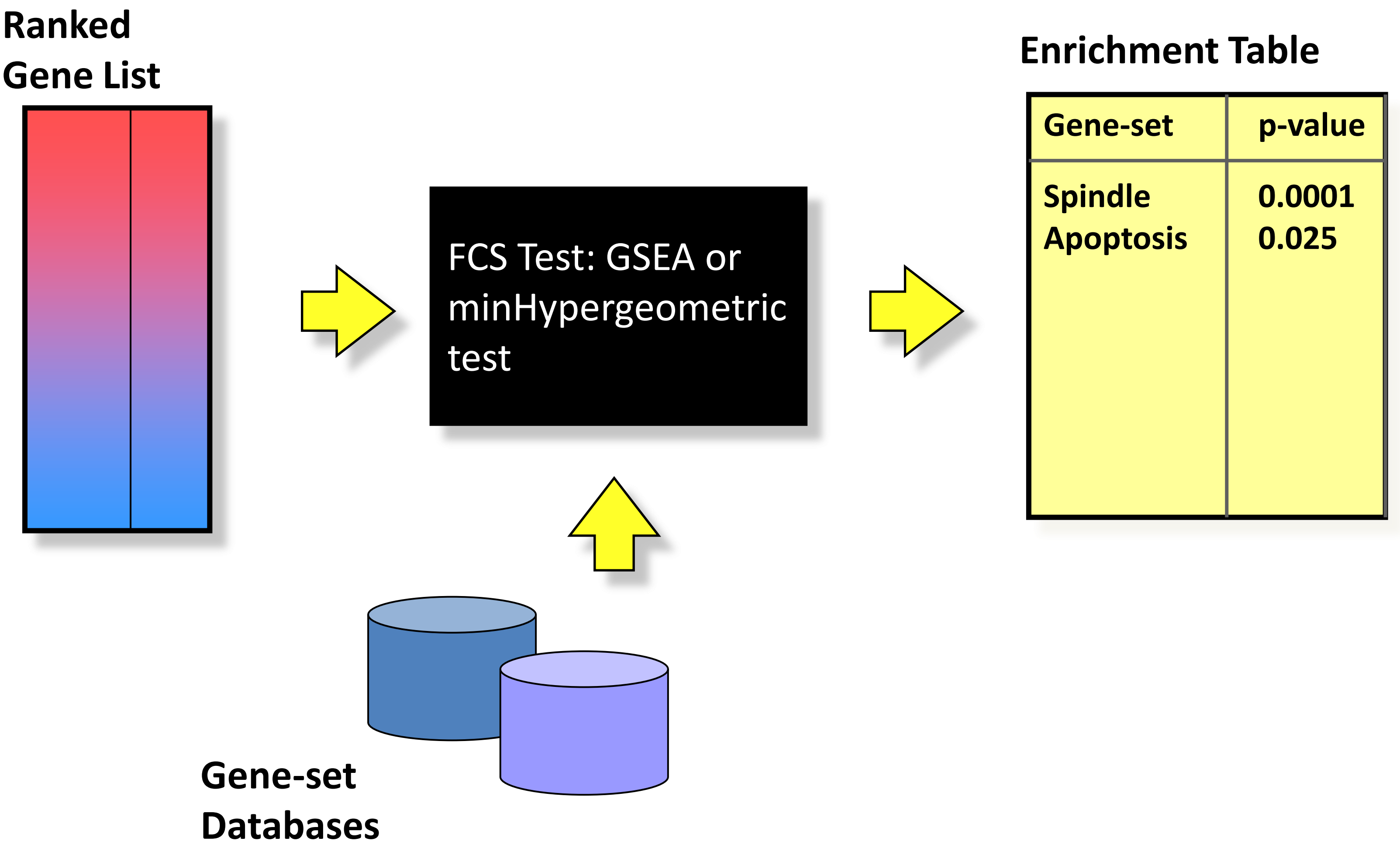
Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

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)



Ranked Gene List

FCS Test: GSEA or minHypergeometric test

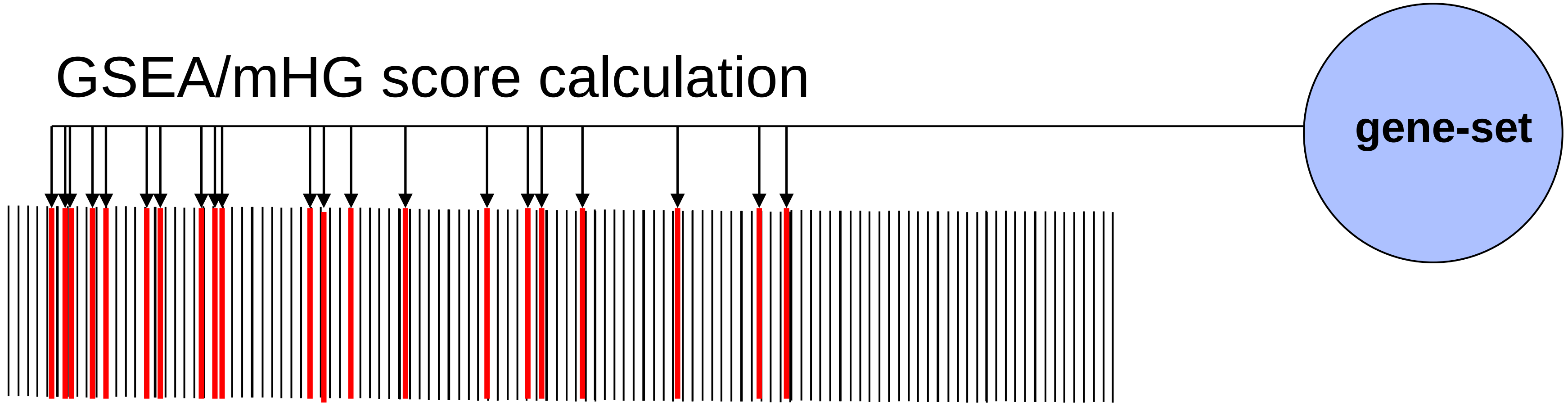
Gene-set Databases

Enrichment Table

Gene-set	p-value
Spindle	0.0001
Apoptosis	0.025

How to score a gene set?

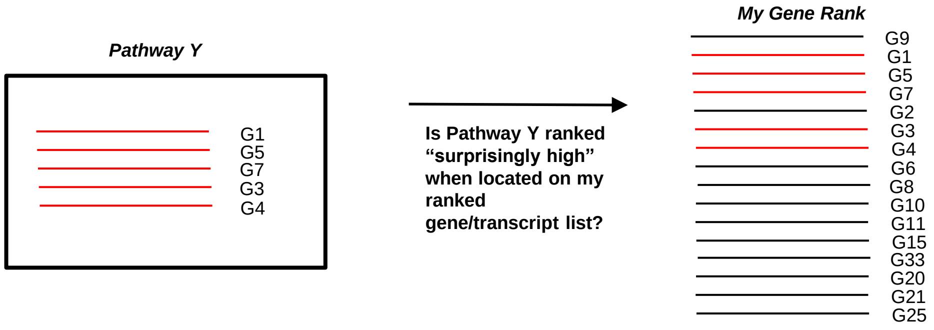
GSEA/mHG score calculation



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

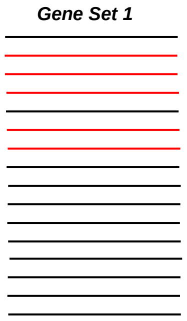
Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 2007 Mar 23;3(3):e39

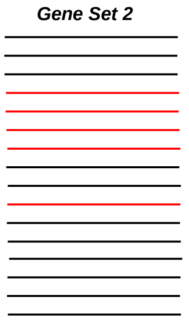
How to score a gene set?



Is Pathway Y ranked "surprisingly high" when located on my ranked gene/transcript list?

Scoring a gene set using the mean rank:





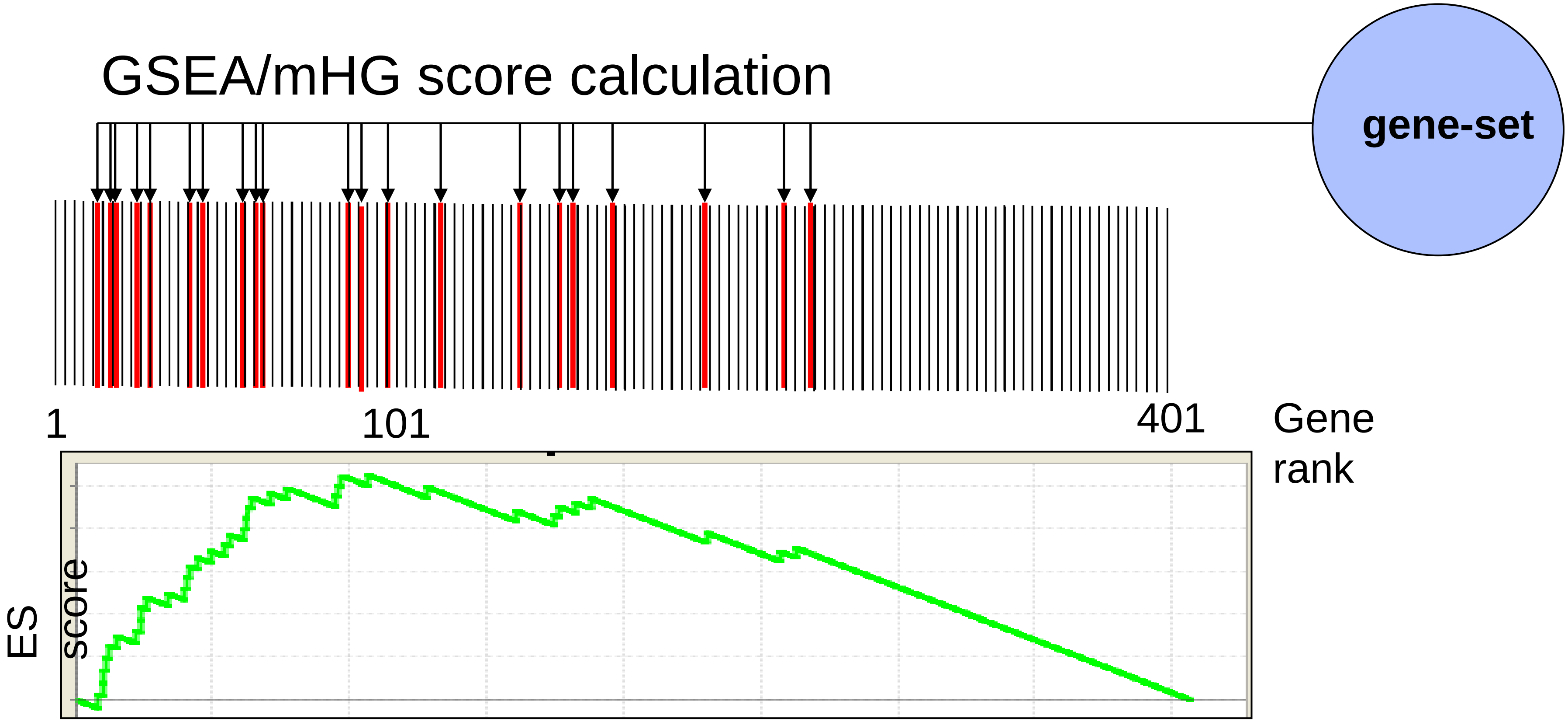
Mean Rank = (2+3+4+6+7) / 5 = 4.4

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

GSEA/mHG score calculation

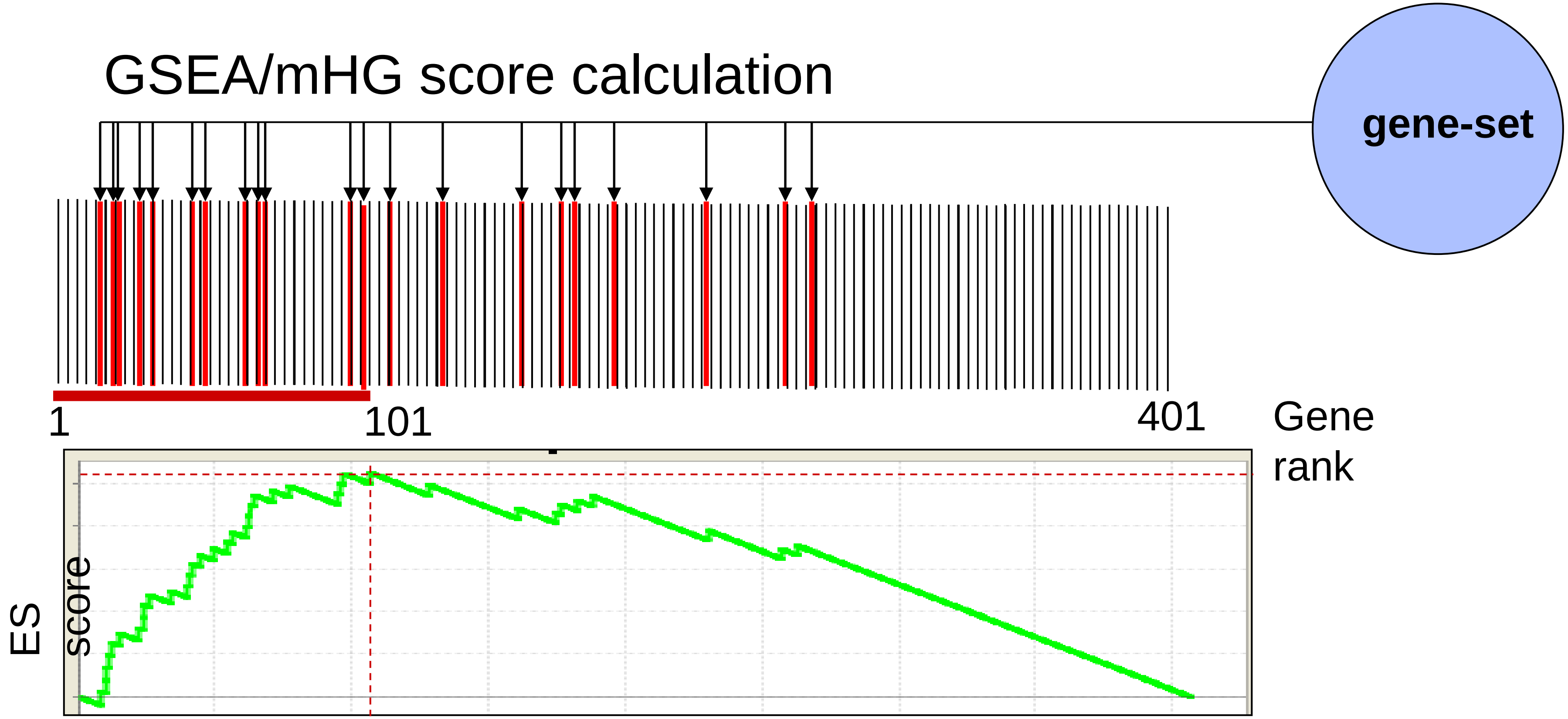


Every present gene (thick red vertical bar) gives a positive contribution,
Every absent gene (black vertical bar) gives a negative contribution

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:

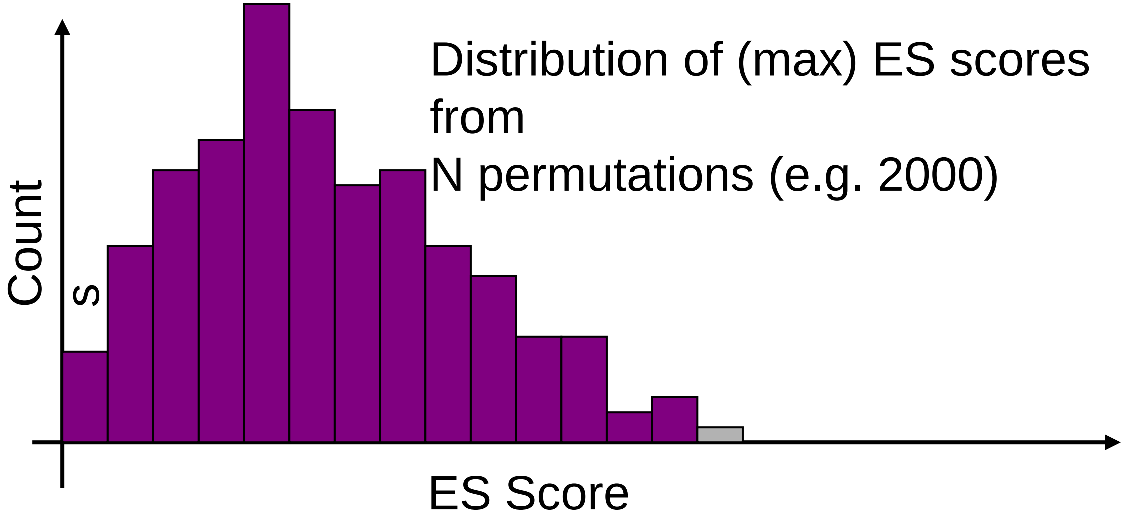
1. Transforming the gene rank into “n” random ranks and then applying the previous procedure in each case. In the end, we will end up with “n” ES values from the random cases.

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)

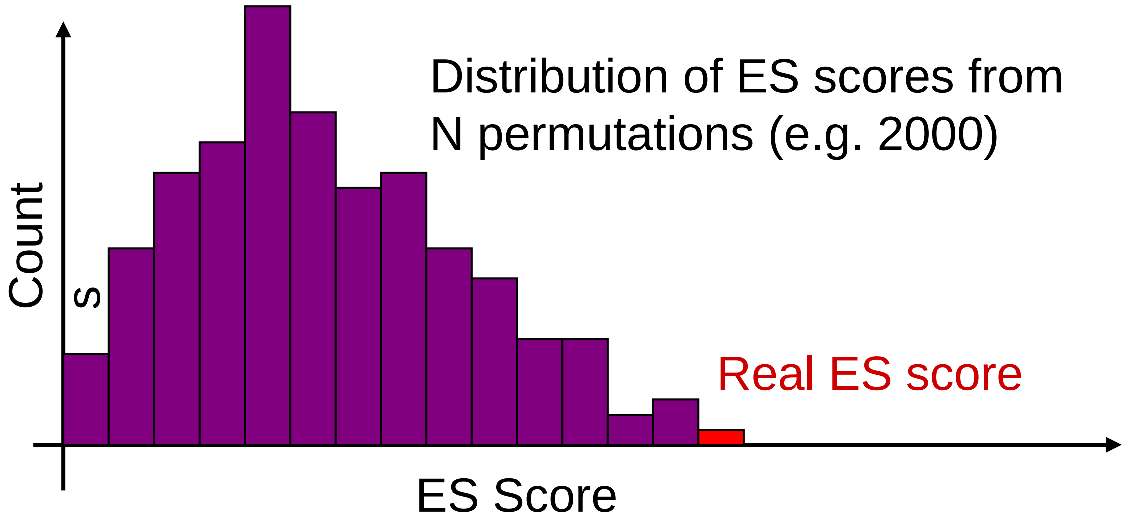
1. Generate null-hypothesis distribution from randomized data



Distribution of (max) ES scores from
N permutations (e.g. 2000)

In statistical terms…

Estimate empirical p-value by comparing observed max ES score to null-hypothesis distribution from randomized data (for every gene-set)



Distribution of ES scores from N permutations (e.g. 2000)

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

Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

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*	0.99
	Translation	

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*	0.99	0.99 x 53/53 = 0.99	0.99
	Translation			

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*	0.99	0.99 x 53/53 = 0.99	0.99
	Translation			

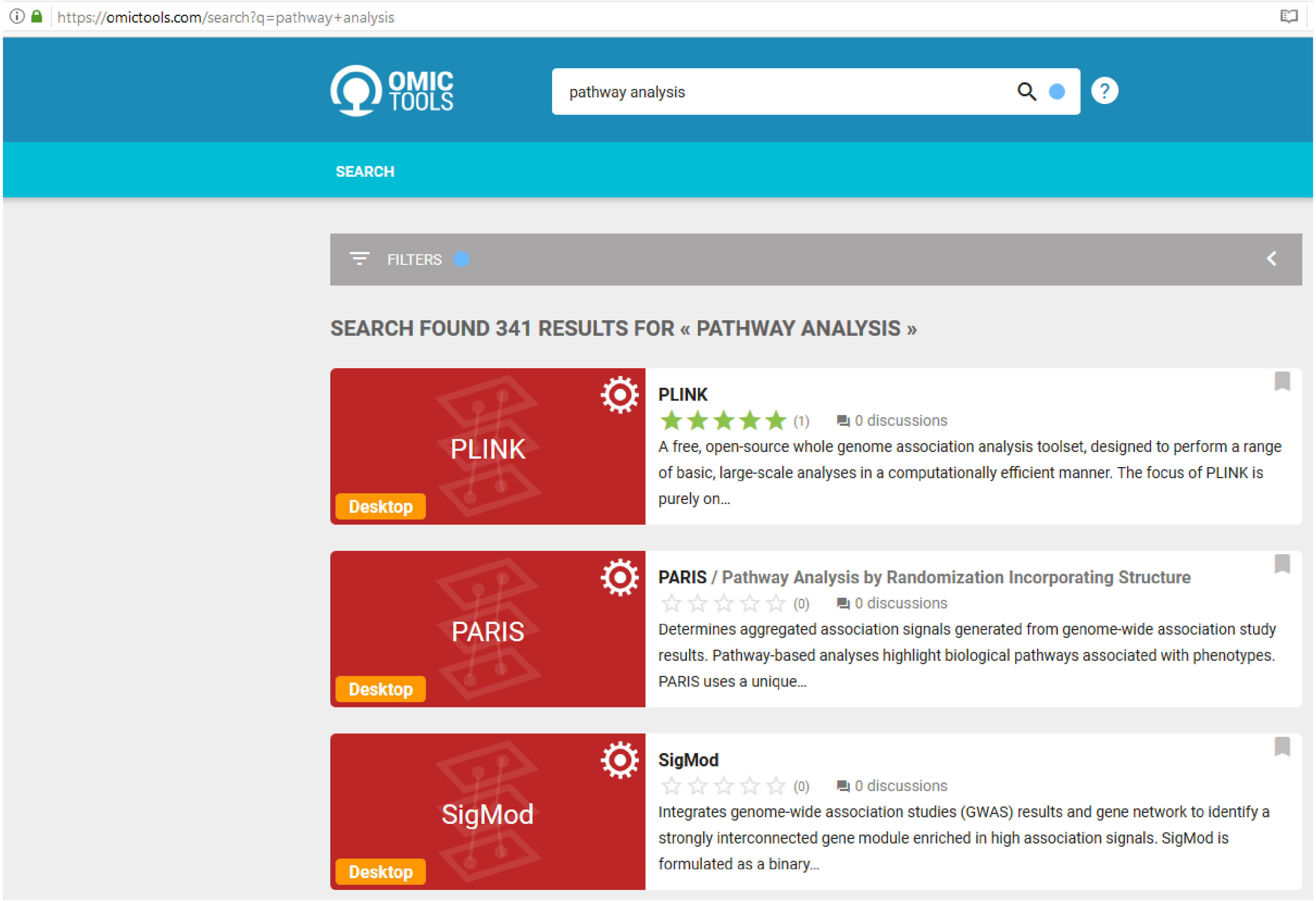
Red: non-significant
Green: significant at FDR < 0.05

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

Contents

2.1. What is Gene Set Analysis.
2.2. Before starting a Gene Set Analysis.
2.3. Gene Set Analysis --ORA
2.4. Gene Set Analysis –FCS
2.5. Multiple testing correction
2.6. Gene Set Analysis --Software

Where to find software?: Omicstools



https://omictools.com/search?q=pathway+analysis

pathway analysis

SEARCH FOUND 341 RESULTS FOR « PATHWAY ANALYSIS »

PLINK

★★★★★ (1) 0 discussions

A free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner. The focus of PLINK is purely on...

PARIS / Pathway Analysis by Randomization Incorporating Structure

☆☆☆☆☆ (0) 0 discussions

Determines aggregated association signals generated from genome-wide association study results. Pathway-based analyses highlight biological pathways associated with phenotypes. PARIS uses a unique...

SigMod

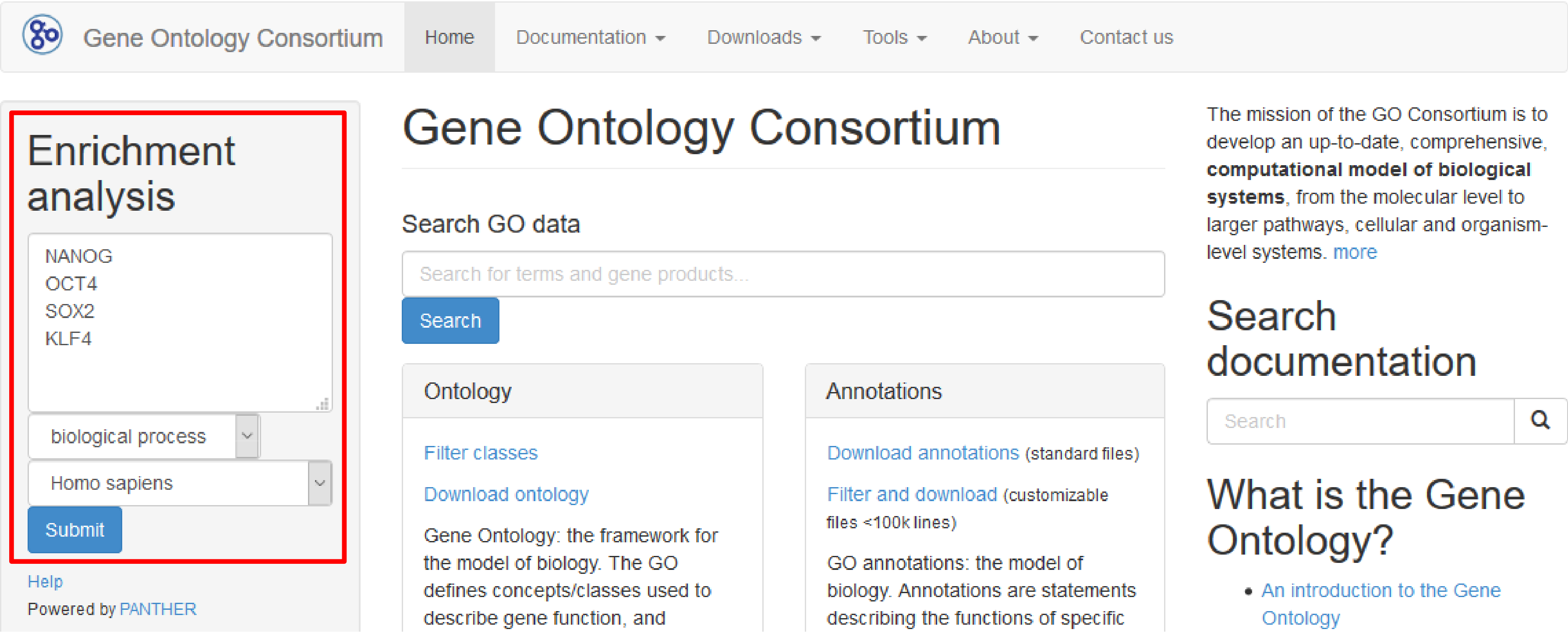
☆☆☆☆☆ (0) 0 discussions

Integrates genome-wide association studies (GWAS) results and gene network to identify a strongly interconnected gene module enriched in high association signals. SigMod is formulated as a binary...

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



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

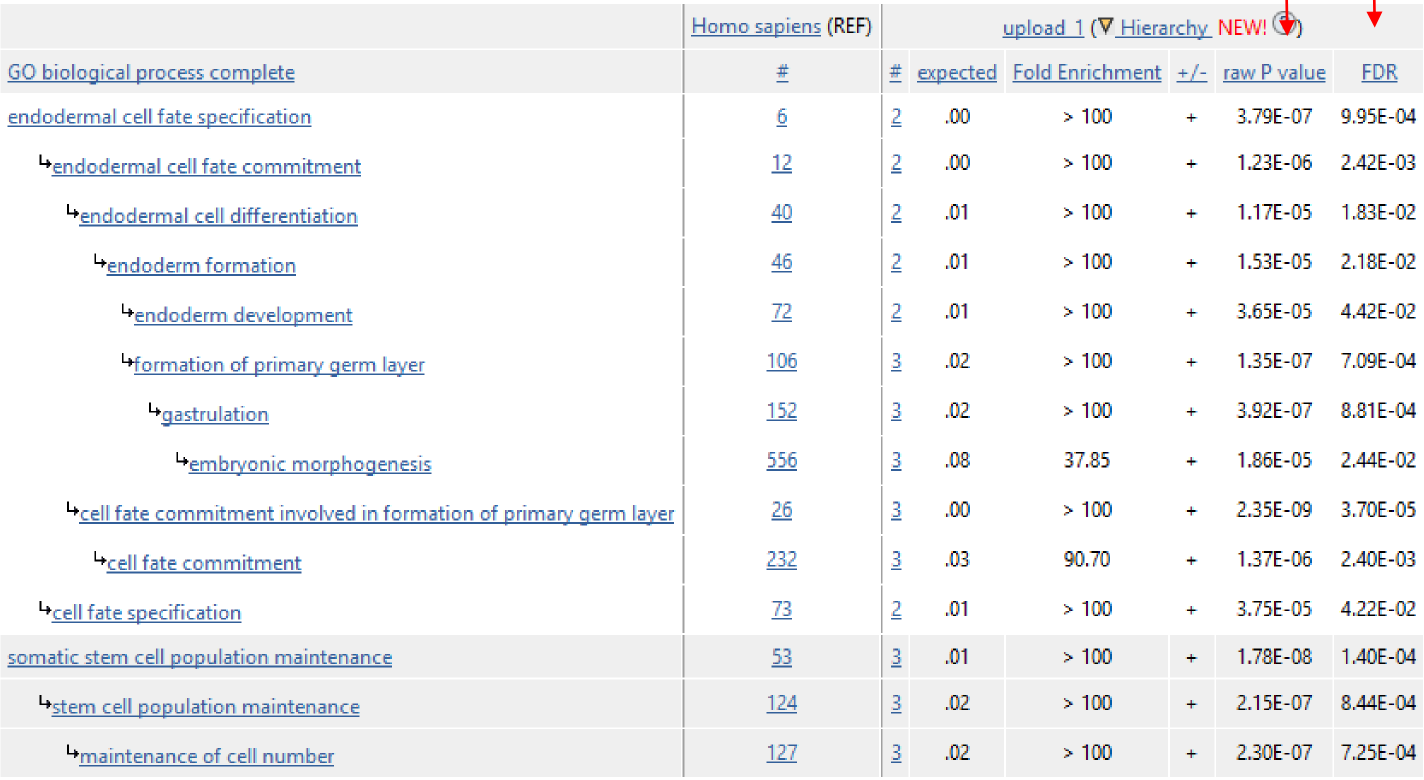
Search documentation

Search

What is the Gene Ontology?

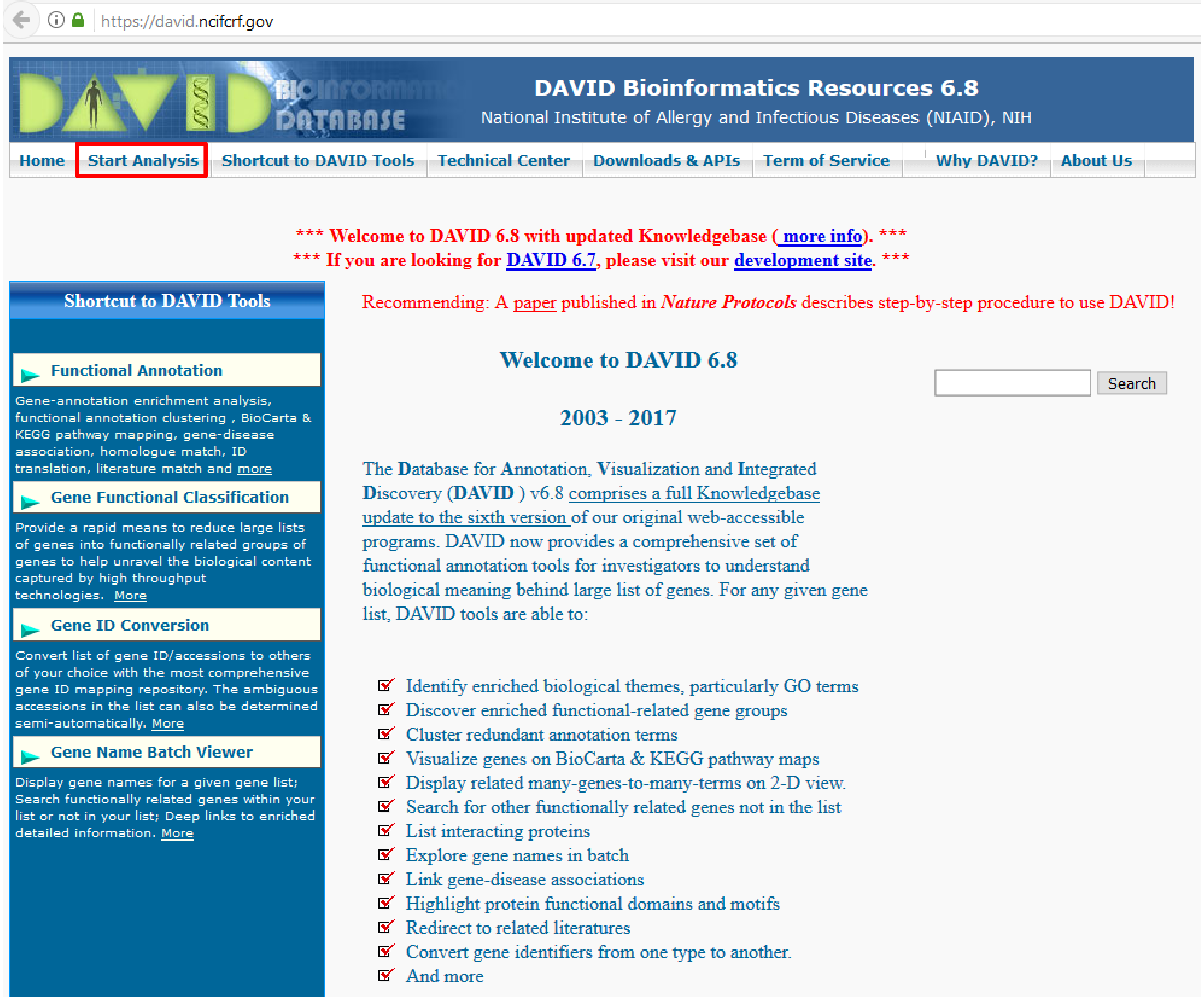
- An introduction to the Gene Ontology

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



DAVID Bioinformatics Resources 6.8

National Institute of Allergy and Infectious Diseases (NIAID), NIH

*** Welcome to DAVID 6.8 with updated Knowledgebase (more info). ***
*** If you are looking for DAVID 6.7, please visit our development site. ***

Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

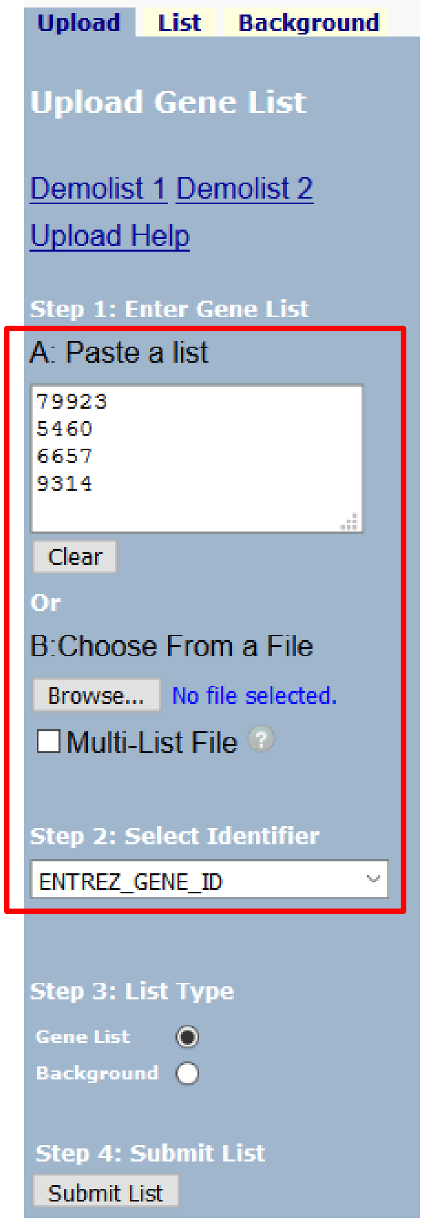
Welcome to DAVID 6.8

2003 - 2017

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- ☑ Identify enriched biological themes, particularly GO terms
- ☑ Discover enriched functional-related gene groups
- ☑ Cluster redundant annotation terms
- ☑ Visualize genes on BioCarta & KEGG pathway maps
- ☑ Display related many-genes-to-many-terms on 2-D view.
- ☑ Search for other functionally related genes not in the list
- ☑ List interacting proteins
- ☑ Explore gene names in batch
- ☑ Link gene-disease associations
- ☑ Highlight protein functional domains and motifs
- ☑ Redirect to related literatures
- ☑ Convert gene identifiers from one type to another.
- ☑ And more

Pathway enrichment analysis software: DAVID



Upload Gene List

Demolist 1 Demolist 2

Upload Help

Step 1: Enter Gene List

A: Paste a list

```
79923
5460
6657
9314
```

Or

B:Choose From a File

No file selected.

☐ Multi-List File

Step 2: Select Identifier

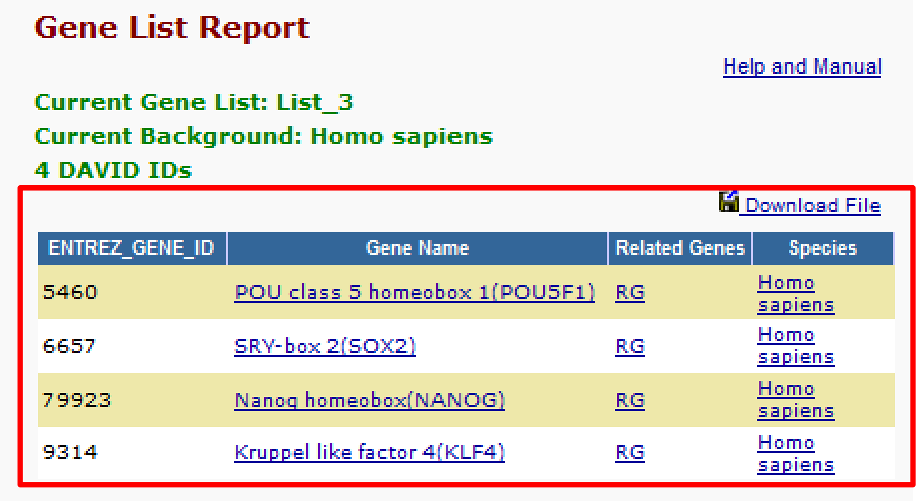
ENTREZ_GENE_ID

Step 3: List Type

Gene List ●

Background ○

Step 4: Submit List



Gene List Report

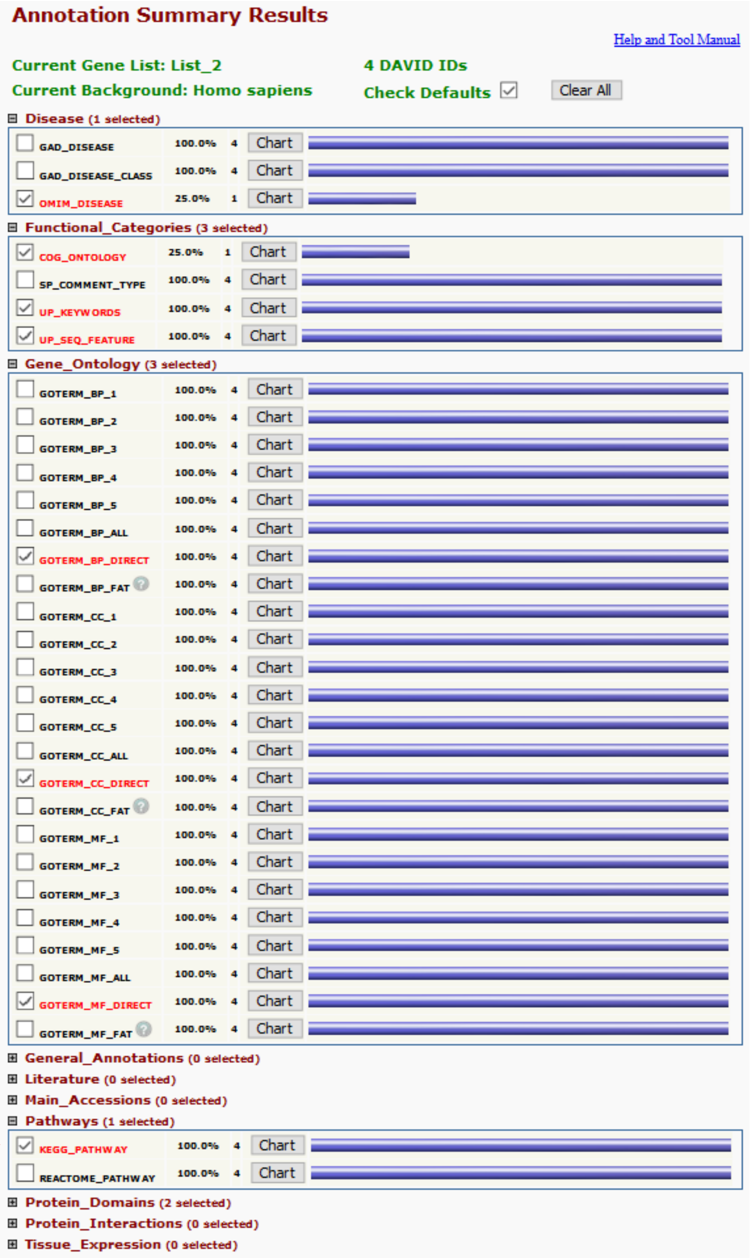
Current Gene List: List_3

Current Background: Homo sapiens

4 DAVID IDs

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



Annotation Summary Results

Help and Tool Manual

Current Gene List: List_2 4 DAVID IDs

Current Background: Homo sapiens Check Defaults ☑ Clear All

⊟ Disease (1 selected)

☐	GAD_DISEASE	100.0%	4	Chart
☐	GAD_DISEASE_CLASS	100.0%	4	Chart
☑	OMIM_DISEASE	25.0%	1	Chart

⊟ Functional_Categories (3 selected)

☑	COG_ONTOLOGY	25.0%	1	Chart
☐	SP_COMMENT_TYPE	100.0%	4	Chart
☑	UP_KEYWORDS	100.0%	4	Chart
☑	UP_SEQ_FEATURE	100.0%	4	Chart

⊟ Gene_Ontology (3 selected)

☐	GOTERM_BP_1	100.0%	4	Chart
☐	GOTERM_BP_2	100.0%	4	Chart
☐	GOTERM_BP_3	100.0%	4	Chart
☐	GOTERM_BP_4	100.0%	4	Chart
☐	GOTERM_BP_5	100.0%	4	Chart
☐	GOTERM_BP_ALL	100.0%	4	Chart
☑	GOTERM_BP_DIRECT	100.0%	4	Chart
☐	GOTERM_BP_FAT	100.0%	4	Chart
☐	GOTERM_CC_1	100.0%	4	Chart
☐	GOTERM_CC_2	100.0%	4	Chart
☐	GOTERM_CC_3	100.0%	4	Chart
☐	GOTERM_CC_4	100.0%	4	Chart
☐	GOTERM_CC_5	100.0%	4	Chart
☐	GOTERM_CC_ALL	100.0%	4	Chart
☑	GOTERM_CC_DIRECT	100.0%	4	Chart
☐	GOTERM_CC_FAT	100.0%	4	Chart
☐	GOTERM_MF_1	100.0%	4	Chart
☐	GOTERM_MF_2	100.0%	4	Chart
☐	GOTERM_MF_3	100.0%	4	Chart
☐	GOTERM_MF_4	100.0%	4	Chart
☐	GOTERM_MF_5	100.0%	4	Chart
☐	GOTERM_MF_ALL	100.0%	4	Chart
☑	GOTERM_MF_DIRECT	100.0%	4	Chart
☐	GOTERM_MF_FAT	100.0%	4	Chart

⊞ General_Annotations (0 selected)

⊞ Literature (0 selected)

⊞ Main_Accessions (0 selected)

⊟ Pathways (1 selected)

☑	KEGG_PATHWAY	100.0%	4	Chart
☐	REACTOME_PATHWAY	100.0%	4	Chart

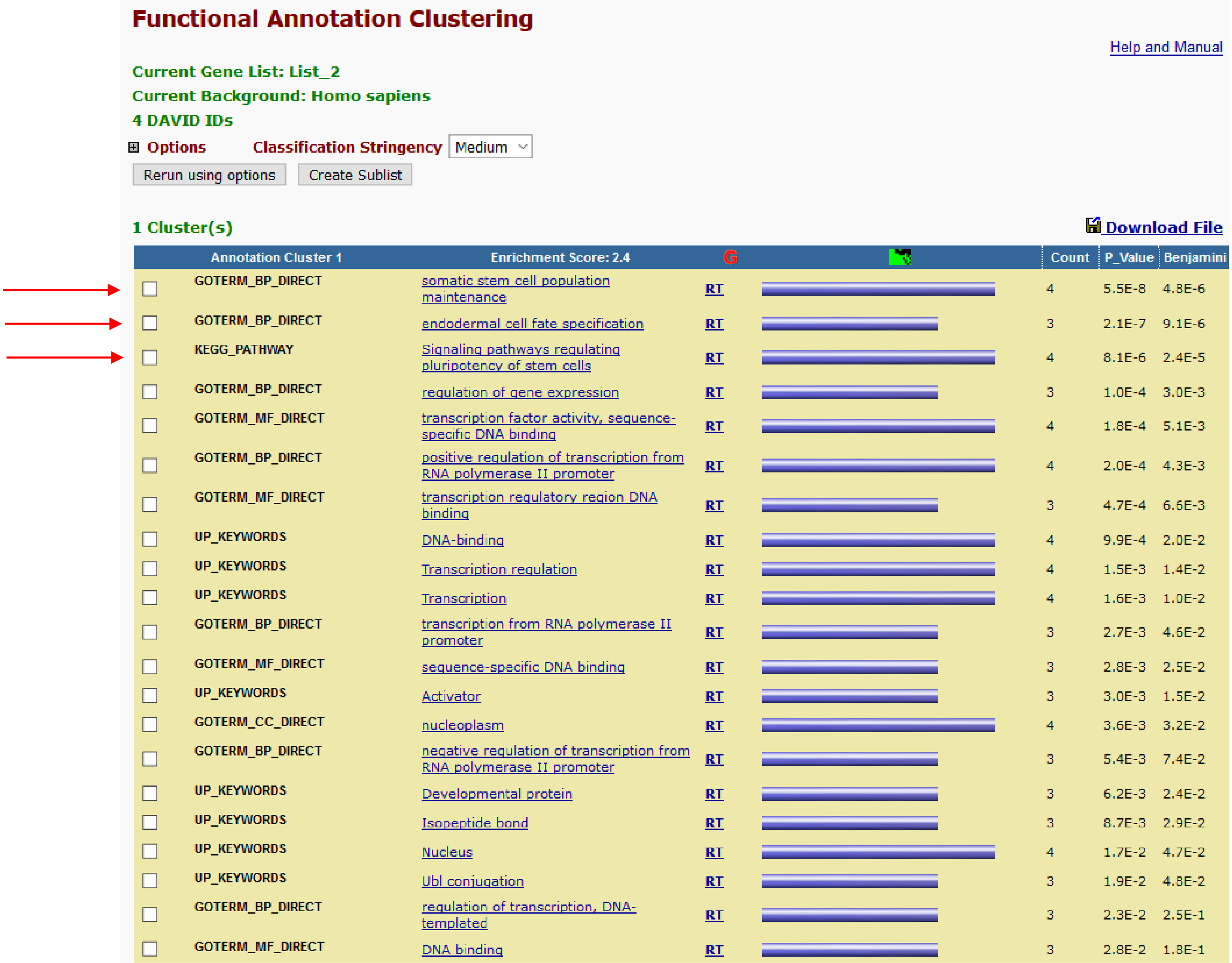
⊞ Protein_Domains (2 selected)

⊞ Protein_Interactions (0 selected)

⊞ Tissue_Expression (0 selected)

Results for KEGG Pathways

Pathway enrichment analysis software: DAVID



Functional Annotation Clustering

Help and Manual

Current Gene List: List_2

Current Background: Homo sapiens

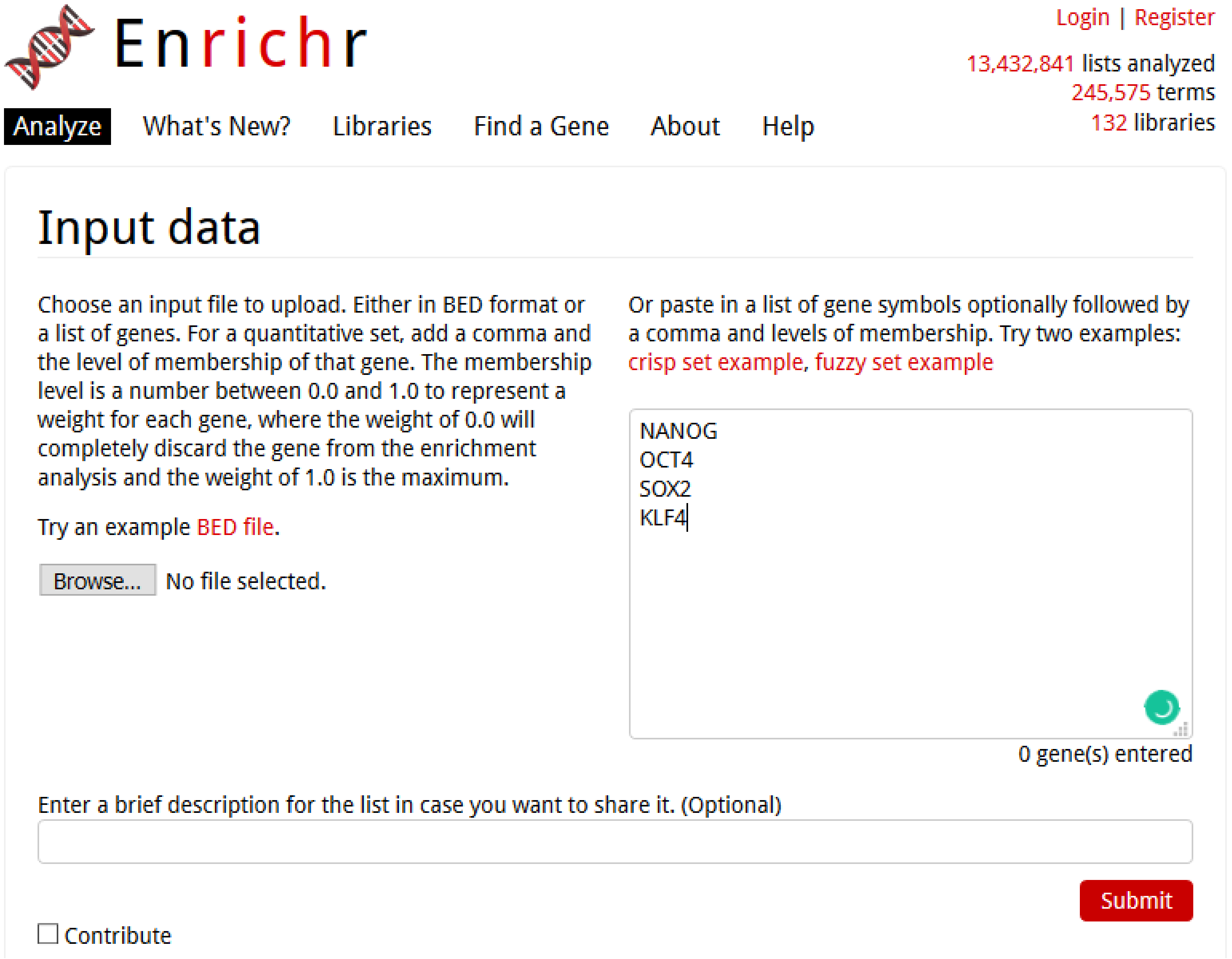
4 DAVID IDs

⊞ **Options** **Classification Stringency** Medium

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



13,432,841 lists analyzed
245,575 terms
132 libraries

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.

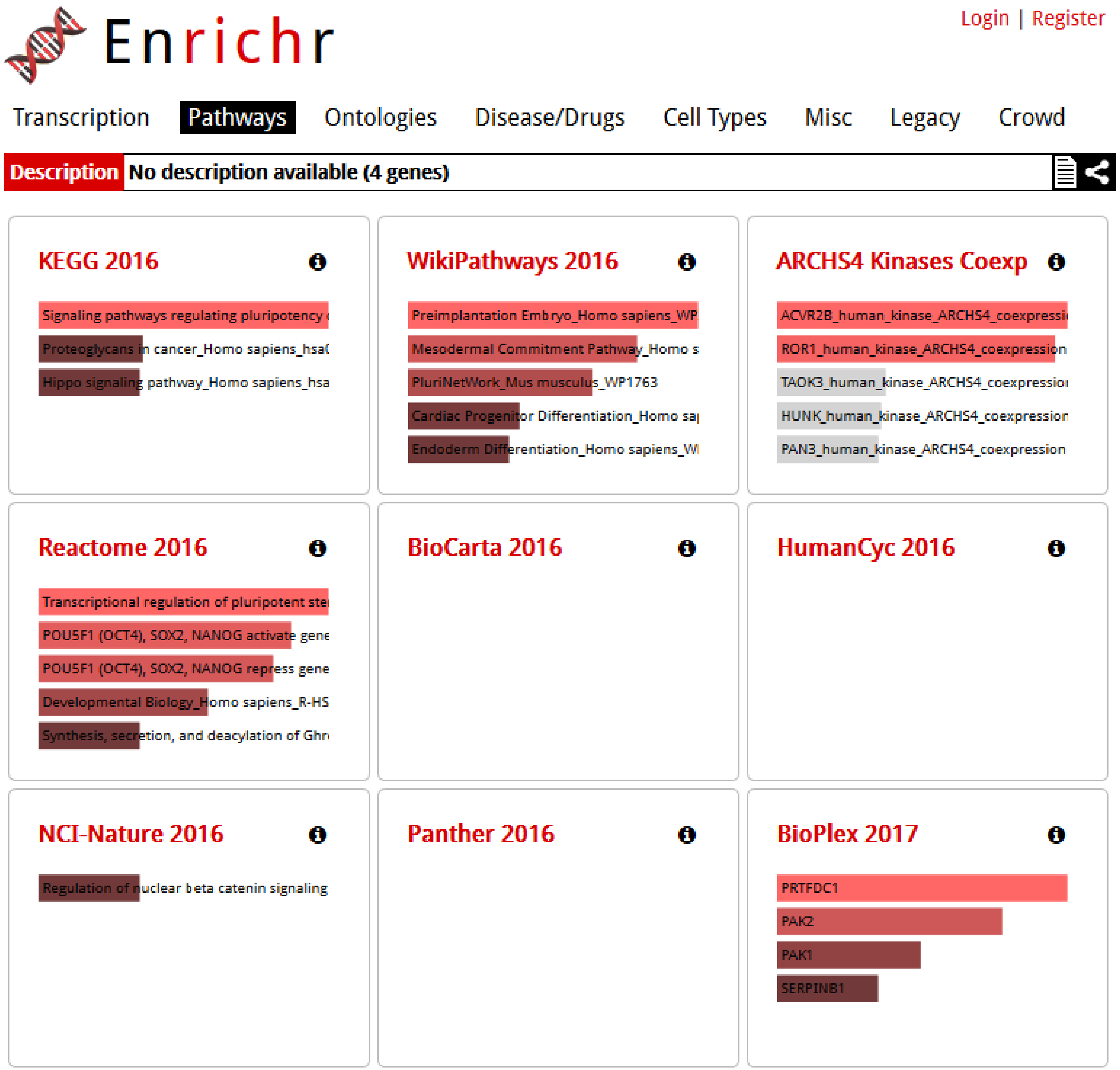
Or paste in a list of gene symbols optionally followed by a comma and levels of membership. Try two examples: crisp set example, fuzzy set example

```
NANOG
OCT4
SOX2
KLF4
```

0 gene(s) entered

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

□ Contribute



Enrichr

Login | Register

Transcription | Pathways | Ontologies | Disease/Drugs | Cell Types | Misc | Legacy | Crowd

Description | No description available (4 genes)

KEGG 2016
- Signaling pathways regulating pluripotency
- Proteoglycans in cancer_Homo sapiens_hsa0
- Hippo signaling pathway_Homo sapiens_hsa

WikiPathways 2016
- Preimplantation Emb ryo_Homo sapiens_WP
- Mesodermal Commitment Pathway_Homo s
- PluriNetWork_Mus musculus_WP1763
- Cardiac Progenitor Differentiation_Homo sa
- Endoderm Differentiation_Homo sapiens_W

ARCHS4 Kinases Coexp
- ACVR2B_human_kinase_ARCHS4_coexpressi
- ROR1_human_kinase_ARCHS4_coexpression
- TAOK3_human_kinase_ARCHS4_coexpressio
- HUNK_human_kinase_ARCHS4_coexpression
- 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

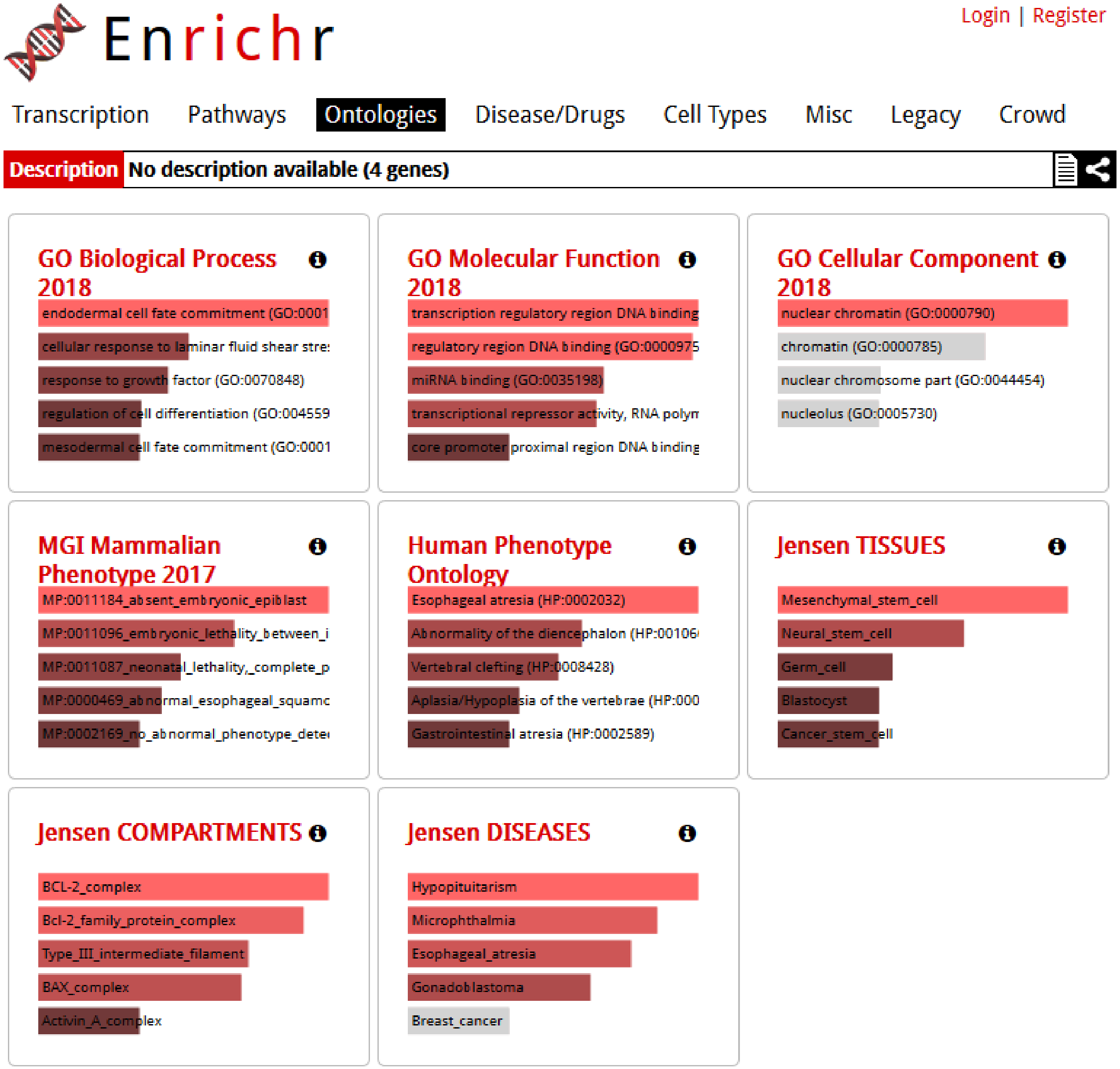
BioCarta 2016

HumanCyc 2016

NCI-Nature 2016
- Regulation of nuclear beta catenin signaling

Panther 2016

BioPlex 2017
- PRTFDC1
- PAK2
- PAK1
- SERPINB1



Enrichr

Login | Register

Transcription Pathways Ontologies Disease/Drugs Cell Types Misc Legacy Crowd

Description No description available (4 genes)

GO Biological Process 2018
- endodermal cell fate commitment (GO:0001
- cellular response to laminar fluid shear stre:
- response to growth factor (GO:0070848)
- regulation of cell differentiation (GO:004559
- mesodermal cell fate commitment (GO:0001

GO Molecular Function 2018
- transcription regulatory region DNA binding
- regulatory region DNA binding (GO:0000975
- miRNA binding (GO:0035198)
- transcriptional repressor activity, RNA polym
- core promoter proximal region DNA binding

GO Cellular Component 2018
- nuclear chromatin (GO:0000790)
- chromatin (GO:0000785)
- nuclear chromosome part (GO:0044454)
- nucleolus (GO:0005730)

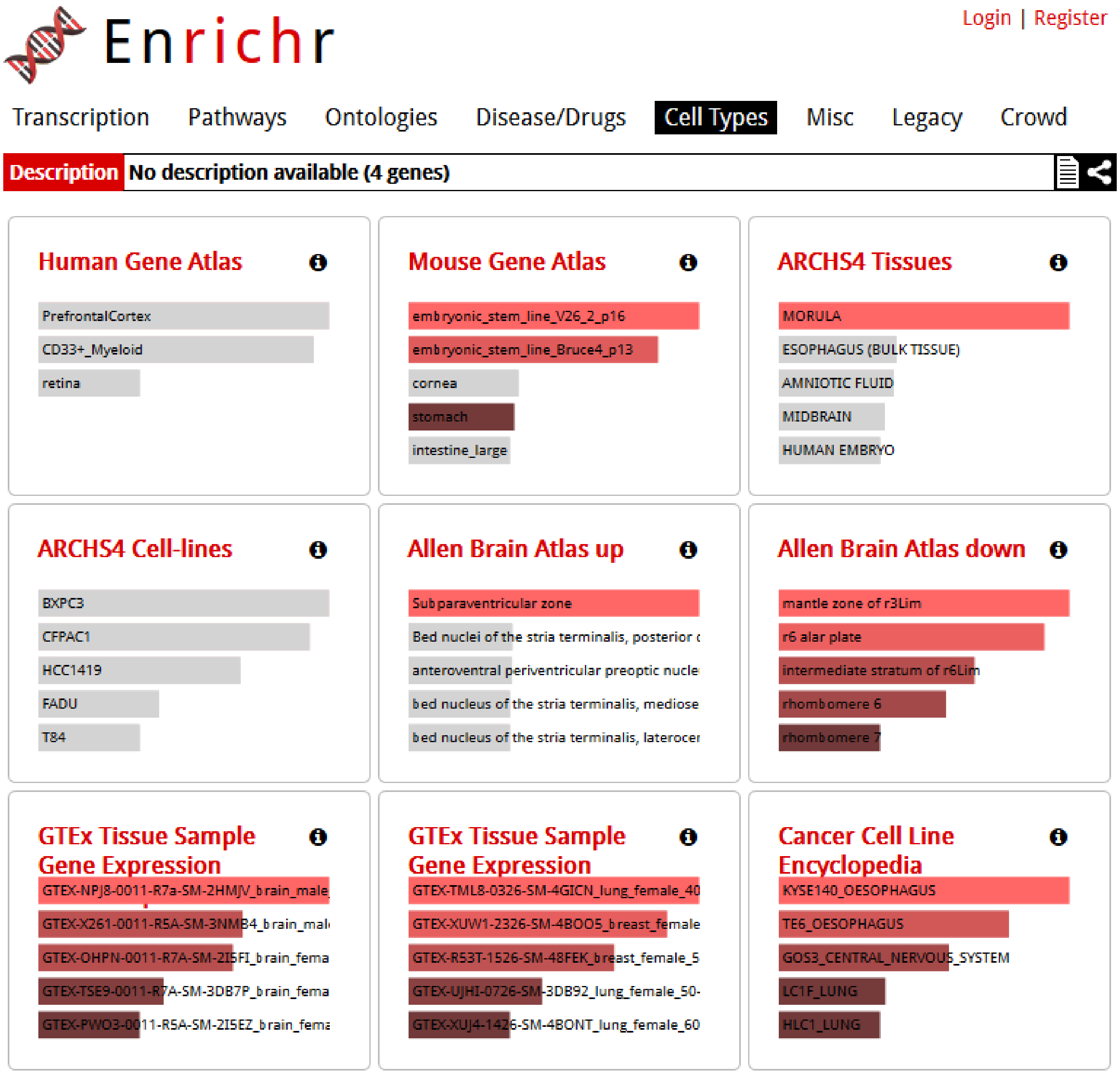
MGI Mammalian Phenotype 2017
- MP:0011184_absent_embryonic_epiblast
- MP:0011096_embryonic_lethality_between_i
- MP:0011087_neonatal_lethality,_complete_p
- MP:0000469_abnormal_esophageal_squamc
- MP:0002169_no_abnormal_phenotype_dete

Human Phenotype Ontology
- Esophageal atresia (HP:0002032)
- Abnormality of the diencephalon (HP:00106
- Vertebral clefting (HP:0008428)
- Aplasia/Hypoplasia of the vertebrae (HP:000
- Gastrointestinal atresia (HP:0002589)

Jensen TISSUES
- Mesenchymal_stem_cell
- Neural_stem_cell
- Germ_cell
- Blastocyst
- Cancer_stem_cell

Jensen COMPARTMENTS
- BCL-2_complex
- Bcl-2_family_protein_complex
- Type_III_intermediate_filament
- BAX_complex
- Activin_A_complex

Jensen DISEASES
- Hypopituitarism
- Microphthalmia
- Esophageal_atresia
- Gonadoblastoma
- Breast_cancer



Enrichr

Login | Register

Transcription Pathways Ontologies Disease/Drugs **Cell Types** Misc Legacy Crowd

Description No description available (4 genes)

Human Gene Atlas

- PrefrontalCortex
- CD33+_Myeloid
- retina

Mouse Gene Atlas

- embryonic_stem_line_V26_2_p16
- embryonic_stem_line_Bruce4_p13
- cornea
- stomach
- intestine_large

ARCHS4 Tissues

- MORULA
- ESOPHAGUS (BULK TISSUE)
- AMNIOTIC FLUID
- MIDBRAIN
- HUMAN EMBRYO

ARCHS4 Cell-lines

- BXPC3
- CFPAC1
- HCC1419
- FADU
- T84

Allen Brain Atlas up

- Subparaventricular zone
- Bed nuclei of the stria terminalis, posterior c
- anteroventral periventricular preoptic nucle
- bed nucleus of the stria terminalis, mediose
- bed nucleus of the stria terminalis, laterocer

Allen Brain Atlas down

- mantle zone of r3Lim
- r6 alar plate
- intermediate stratum of r6Lim
- rhombomere 6
- rhombomere 7

GTEx Tissue Sample Gene Expression

- GTEX-NPJ8-0011-R7a-SM-2HMJV_brain_male_
- GTEX-X261-0011-R5A-SM-3NMB4_brain_mal
- GTEX-OHPN-0011-R7A-SM-2I5FI_brain_fema
- GTEX-TSE9-0011-R7A-SM-3DB7P_brain_fema
- GTEX-PWO3-0011-R5A-SM-2I5EZ_brain_fema

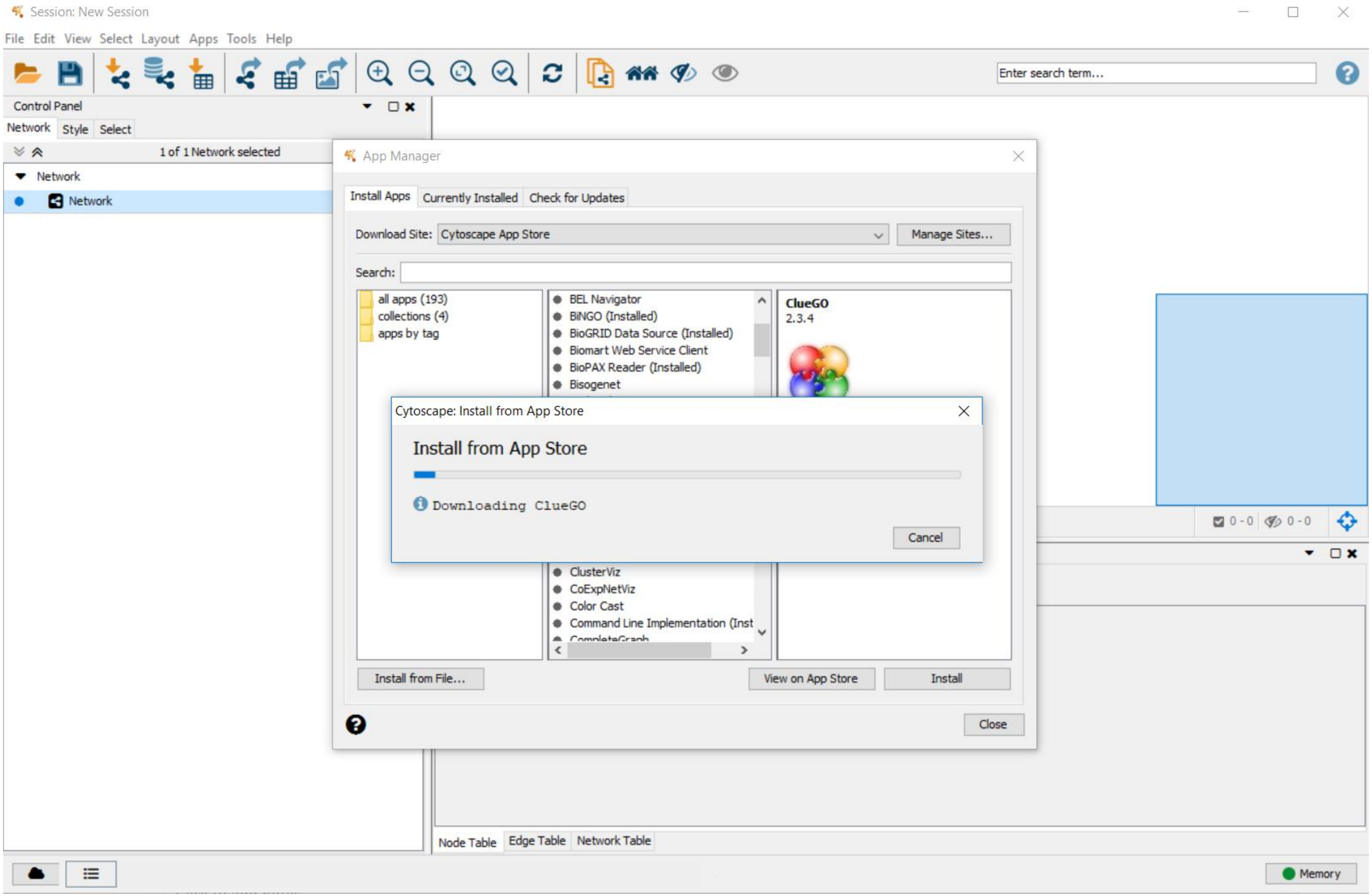
GTEx Tissue Sample Gene Expression

- GTEX-TML8-0326-SM-4GICN_lung_female_40
- GTEX-XUW1-2326-SM-4BOO5_breast_female
- GTEX-R53T-1526-SM-48FEK_breast_female_5
- GTEX-UJHI-0726-SM-3DB92_lung_female_50-
- GTEX-XUJ4-1426-SM-4BONT_lung_female_60

Cancer Cell Line Encyclopedia

- KYSE140_OESOPHAGUS
- TE6_OESOPHAGUS
- GOS3_CENTRAL_NERVOUS_SYSTEM
- LC1F_LUNG
- HLC1_LUNG

Pathway enrichment analysis software: Cytoscape / ClueGO



Session: New Session

Control Panel

1 of 1 Network selected

Network
- Network

App Manager

Download Site: Cytoscape App Store

Search:

- all apps (193)
- collections (4)
- apps by tag

- BEL Navigator
- BiNGO (Installed)
- BioGRID Data Source (Installed)
- Biomart Web Service Client
- BioPAX Reader (Installed)
- Bisogenet
- ClusterViz
- CoExpNetViz
- Color Cast
- Command Line Implementation (Inst
- CompleteGraph

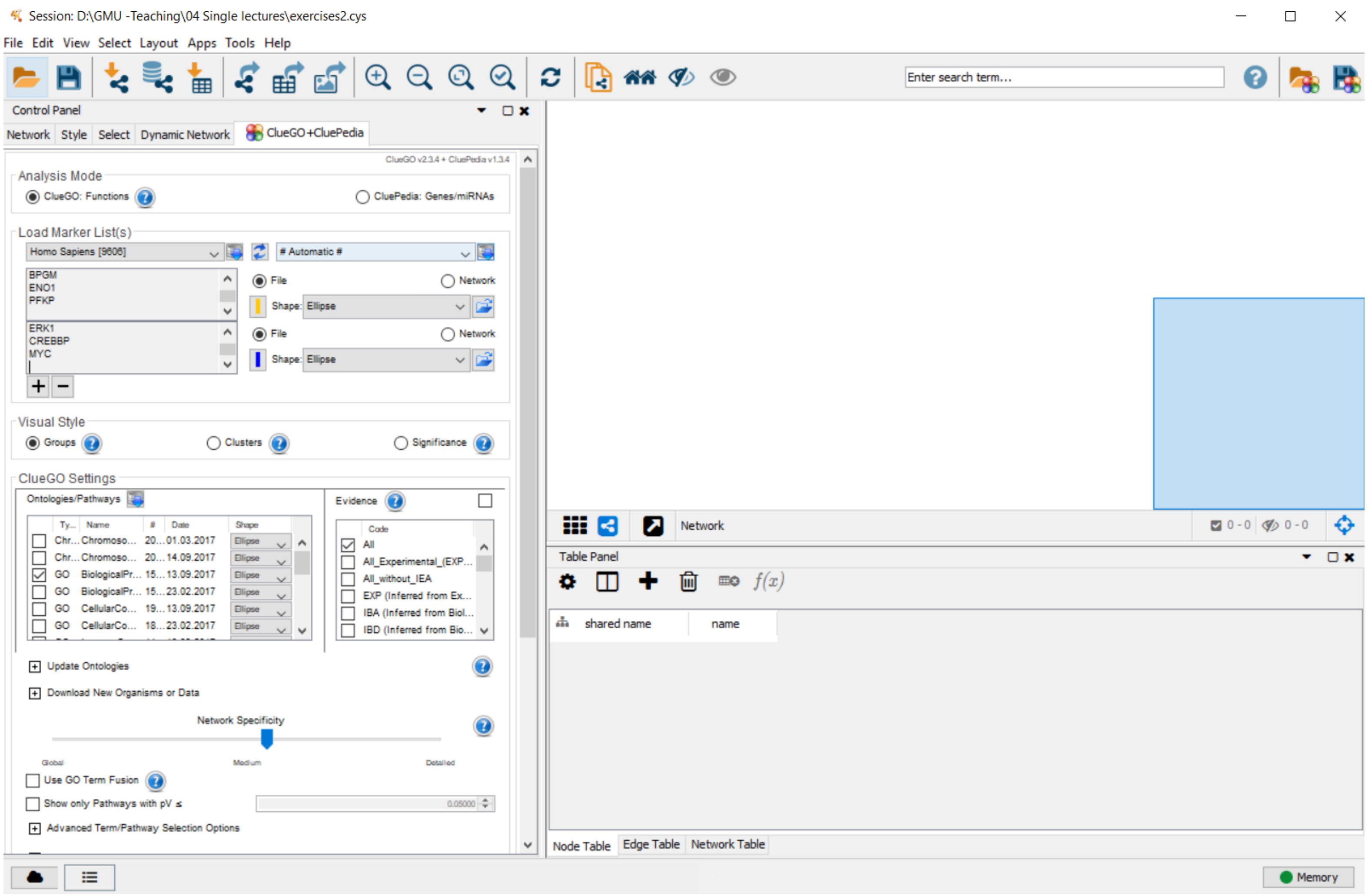
ClueGO
2.3.4

Cytoscape: Install from App Store

Install from App Store

Downloading ClueGO

Pathway enrichment analysis software: Cytoscape / ClueGO



File Edit View Select Layout Apps Tools Help
Enter search term...
Control Panel
Network Style Select Dynamic Network ClueGO+CluePedia
ClueGO v2.3.4 + CluePedia v1.3.4
Network
Table Panel
f(x)
Node Table Edge Table Network Table
Memory

Session: D:\GMU -Teaching\04 Single lectures\exercises2.cys

Analysis Mode
● ClueGO: Functions ○ CluePedia: Genes/miRNAs

Load Marker List(s)
Homo Sapiens [9606] — # Automatic #

BPGM
ENO1
PFKP

● File ○ Network — Shape: Ellipse

ERK1
CREBBP
MYC

● File ○ Network — Shape: Ellipse

Visual Style
● Groups ○ Clusters ○ Significance

ClueGO Settings

Ontologies/Pathways

	Ty...	Name	#	Date	Shape
□	Chr...	Chromoso...	20...	01.03.2017	Ellipse
□	Chr...	Chromoso...	20...	14.09.2017	Ellipse
☑	GO	BiologicalPr...	15...	13.09.2017	Ellipse
□	GO	BiologicalPr...	15...	23.02.2017	Ellipse
□	GO	CellularCo...	19...	13.09.2017	Ellipse
□	GO	CellularCo...	18...	23.02.2017	Ellipse

Evidence □

	Code
☑	All
□	All_Experimental_(EXP...
□	All_without_IEA
□	EXP (Inferred from Ex...
□	IBA (Inferred from Biol...
□	IBD (Inferred from Bio...

⊞ Update Ontologies

⊞ Download New Organisms or Data

Network Specificity: Global — Medium — Detailed

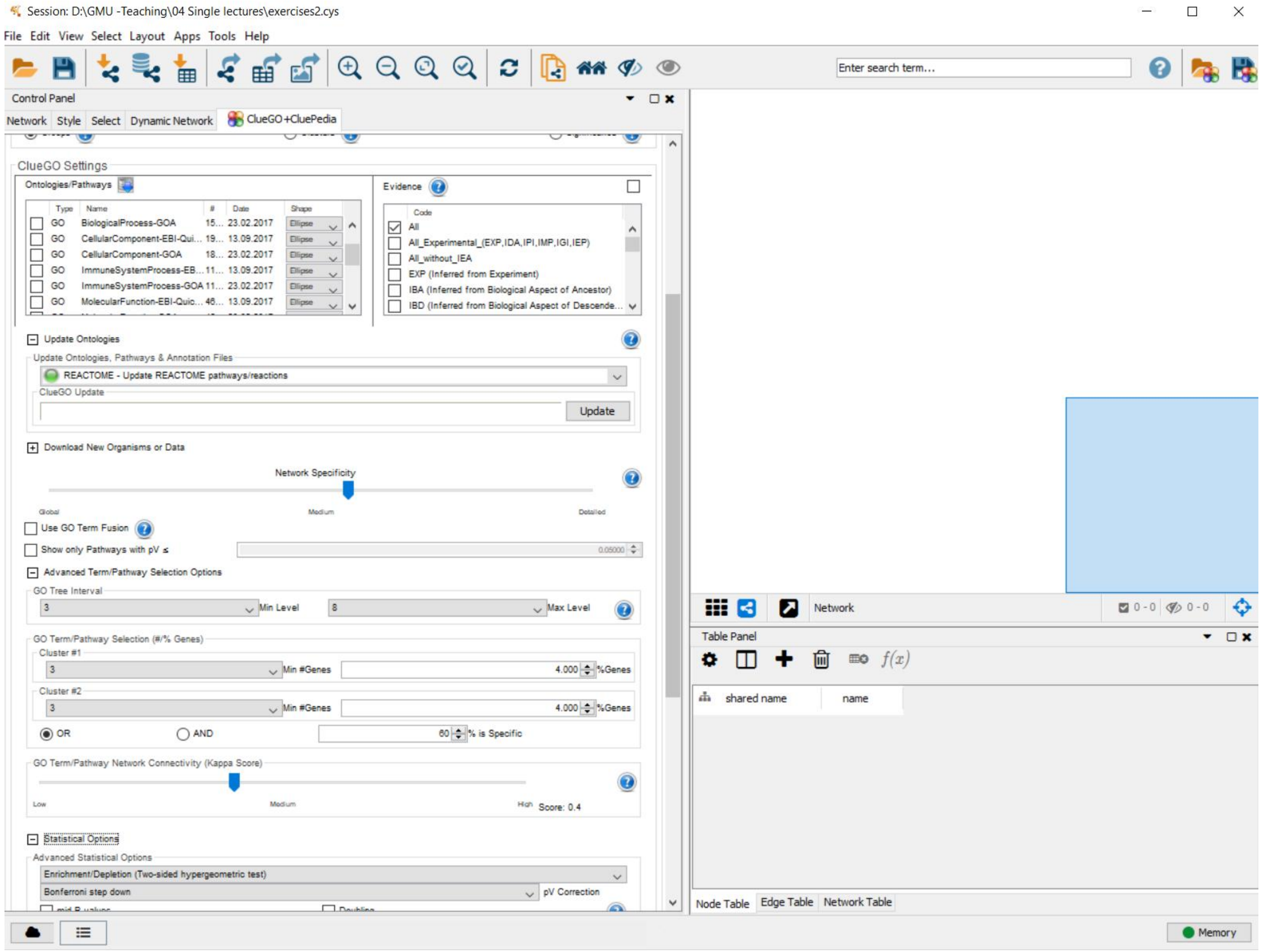
□ Use GO Term Fusion

□ Show only Pathways with pV ≤ 0.05000

⊞ Advanced Term/Pathway Selection Options

shared name	name

Pathway enrichment analysis software: Cytoscape / ClueGO


Session: D:\GMU -Teaching\04 Single lectures\exercises2.cys
File Edit View Select Layout Apps Tools Help
Enter search term...
Control Panel
Network Style Select Dynamic Network ClueGO+CluePedia
Network
Table Panel
Node Table Edge Table Network Table
Memory

ClueGO Settings

Ontologies/Pathways

	Type	Name	#	Date	Shape
□	GO	BiologicalProcess-GOA	15...	23.02.2017	Ellipse
□	GO	CellularComponent-EBI-Qui...	19...	13.09.2017	Ellipse
□	GO	CellularComponent-GOA	18...	23.02.2017	Ellipse
□	GO	ImmuneSystemProcess-EB...	11...	13.09.2017	Ellipse
□	GO	ImmuneSystemProcess-GOA	11...	23.02.2017	Ellipse
□	GO	MolecularFunction-EBI-Quic...	46...	13.09.2017	Ellipse

Evidence □

	Code
☑	All
□	All_Experimental_(EXP,IDA,IPI,IMP,IGI,IEP)
□	All_without_IEA
□	EXP (Inferred from Experiment)
□	IBA (Inferred from Biological Aspect of Ancestor)
□	IBD (Inferred from Biological Aspect of Descende...

⊟ Update Ontologies

Update Ontologies, Pathways & Annotation Files: REACTOME - Update REACTOME pathways/reactions

ClueGO Update: Update

⊞ Download New Organisms or Data

Network Specificity: Global — Medium — Detailed

□ Use GO Term Fusion

□ Show only Pathways with pV ≤ 0.05000

⊟ Advanced Term/Pathway Selection Options

GO Tree Interval: 3 Min Level, 8 Max Level

GO Term/Pathway Selection (#/% Genes)

Cluster #1: 3 Min #Genes, 4.000 %Genes

Cluster #2: 3 Min #Genes, 4.000 %Genes

● OR ○ AND 60 % is Specific

GO Term/Pathway Network Connectivity (Kappa Score): Low — Medium — High Score: 0.4

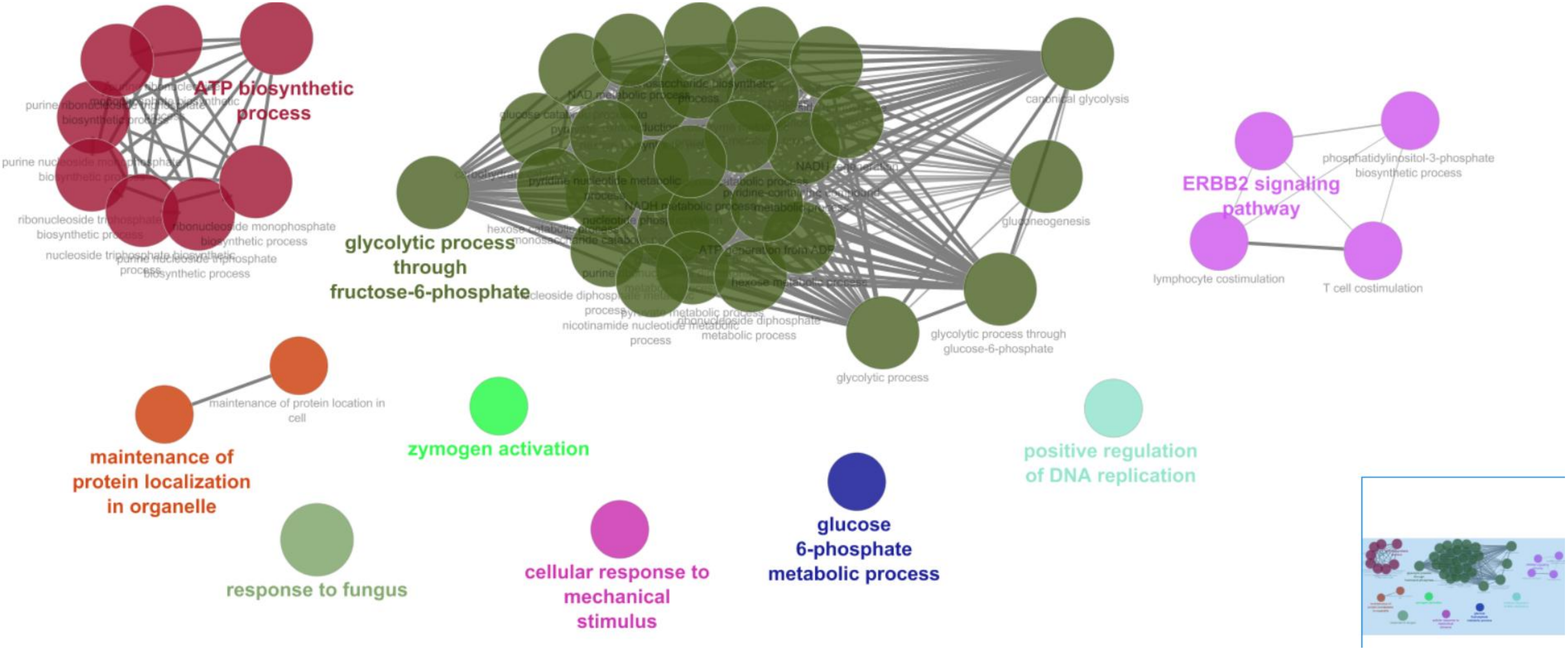
⊟ Statistical Options

Advanced Statistical Options

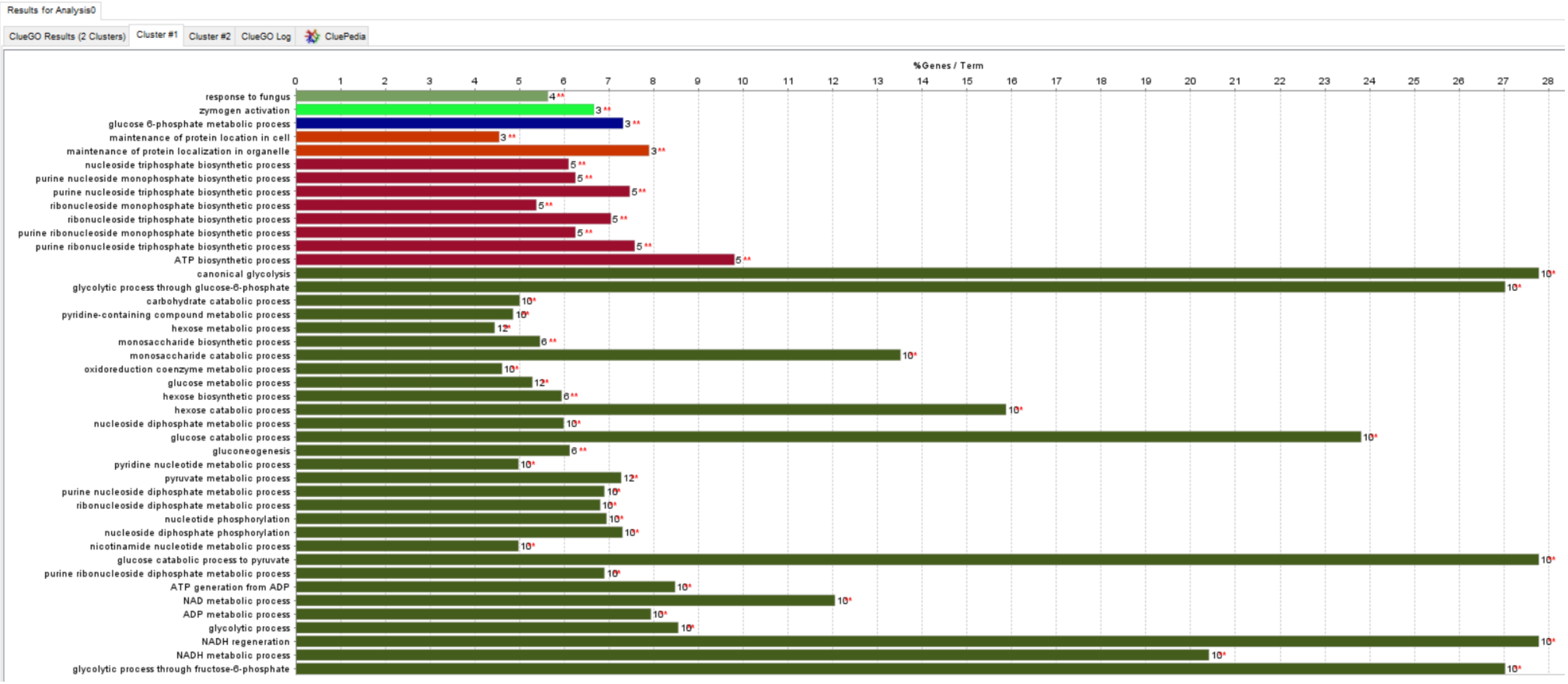
Enrichment/Depletion (Two-sided hypergeometric test)

Bonferroni step down pV Correction

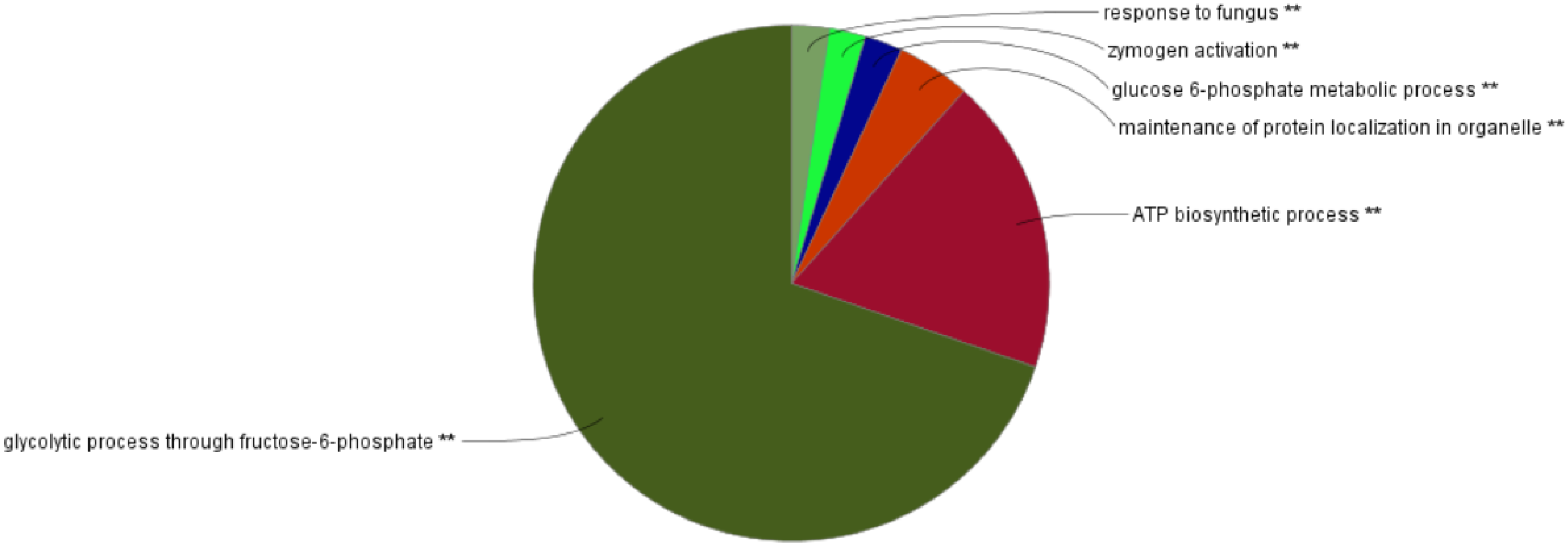
Pathway enrichment analysis software: Cytoscape / ClueGO


ATP biosynthetic process
purine ribonucleoside triphosphate biosynthetic process
purine nucleoside monophosphate biosynthetic process
ribonucleoside triphosphate biosynthetic process
ribonucleoside monophosphate biosynthetic process
nucleoside triphosphate biosynthetic process
purine nucleoside triphosphate biosynthetic process
glycolytic process through fructose-6-phosphate
NAD metabolic process
monosaccharide biosynthetic process
glucose catabolic process to pyruvate
pyridine nucleotide metabolic process
NADH metabolic process
NADH regeneration
pyridine-containing compound metabolic process
nucleotide phosphorylation
hexose catabolic process
monosaccharide catabolic process
ATP generation from ADP
hexose metabolic process
nucleoside diphosphate metabolic process
pyruvate metabolic process
ribonucleoside diphosphate metabolic process
nicotinamide nucleotide metabolic process
canonical glycolysis
gluconeogenesis
glycolytic process through glucose-6-phosphate
glycolytic process
ERBB2 signaling pathway
phosphatidylinositol-3-phosphate biosynthetic process
lymphocyte costimulation
T cell costimulation
maintenance of protein location in cell
maintenance of protein localization in organelle
zymogen activation
positive regulation of DNA replication
response to fungus
cellular response to mechanical stimulus
glucose 6-phosphate metabolic process

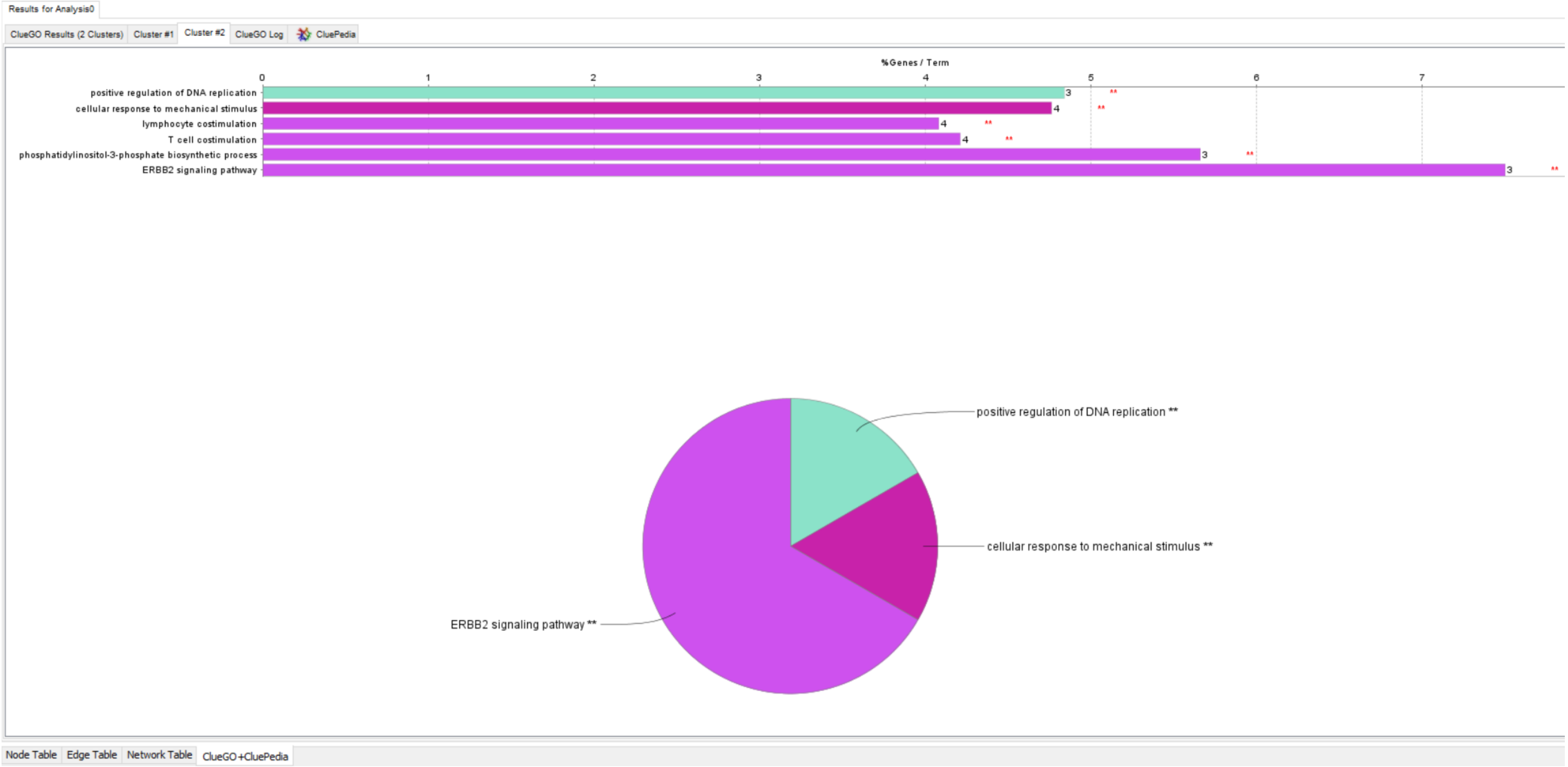
Pathway enrichment analysis software: Cytoscape / ClueGO


Results for Analysis0
ClueGO Results (2 Clusters)
Cluster #1
Cluster #2
ClueGO Log
CluePedia
%Genes / Term
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
response to fungus 4**
zymogen activation 3**
glucose 6-phosphate metabolic process 3**
maintenance of protein location in cell 3**
maintenance of protein localization in organelle 3**
nucleoside triphosphate biosynthetic process 5**
purine nucleoside monophosphate biosynthetic process 5**
purine nucleoside triphosphate biosynthetic process 5**
ribonucleoside monophosphate biosynthetic process 5**
ribonucleoside triphosphate biosynthetic process 5**
purine ribonucleoside monophosphate biosynthetic process 5**
purine ribonucleoside triphosphate biosynthetic process 5**
ATP biosynthetic process 5**
canonical glycolysis 10*
glycolytic process through glucose-6-phosphate 10*
carbohydrate catabolic process 10*
pyridine-containing compound metabolic process 10*
hexose metabolic process 12*
monosaccharide biosynthetic process 6**
monosaccharide catabolic process 10*
oxidoreduction coenzyme metabolic process 10*
glucose metabolic process 12*
hexose biosynthetic process 6**
hexose catabolic process 10*
nucleoside diphosphate metabolic process 10*
glucose catabolic process 10*
gluconeogenesis 6**
pyridine nucleotide metabolic process 10*
pyruvate metabolic process 12*
purine nucleoside diphosphate metabolic process 10*
ribonucleoside diphosphate metabolic process 10*
nucleotide phosphorylation 10*
nucleoside diphosphate phosphorylation 10*
nicotinamide nucleotide metabolic process 10*
glucose catabolic process to pyruvate 10*
purine ribonucleoside diphosphate metabolic process 10*
ATP generation from ADP 10*
NAD metabolic process 10*
ADP metabolic process 10*
glycolytic process 10*
NADH regeneration 10*
NADH metabolic process 10*
glycolytic process through fructose-6-phosphate 10*

Pathway enrichment analysis software: Cytoscape / ClueGO



Pathway enrichment analysis software: Cytoscape / ClueGO


Results for Analysis0
ClueGO Results (2 Clusters)
Cluster #1
Cluster #2
ClueGO Log
CluePedia
%Genes / Term
0
1
2
3
4
5
6
7
positive regulation of DNA replication
3 **
cellular response to mechanical stimulus
4 **
lymphocyte costimulation
4 **
T cell costimulation
4 **
phosphatidylinositol-3-phosphate biosynthetic process
3 **
ERBB2 signaling pathway
3 **
positive regulation of DNA replication **
cellular response to mechanical stimulus **
ERBB2 signaling pathway **
Node Table
Edge Table
Network Table
ClueGO+CluePedia

Enrichment Map

GENE SETS

Spindle

Gene.A
Gene.B
Gene.C

Ca^{++} Channels

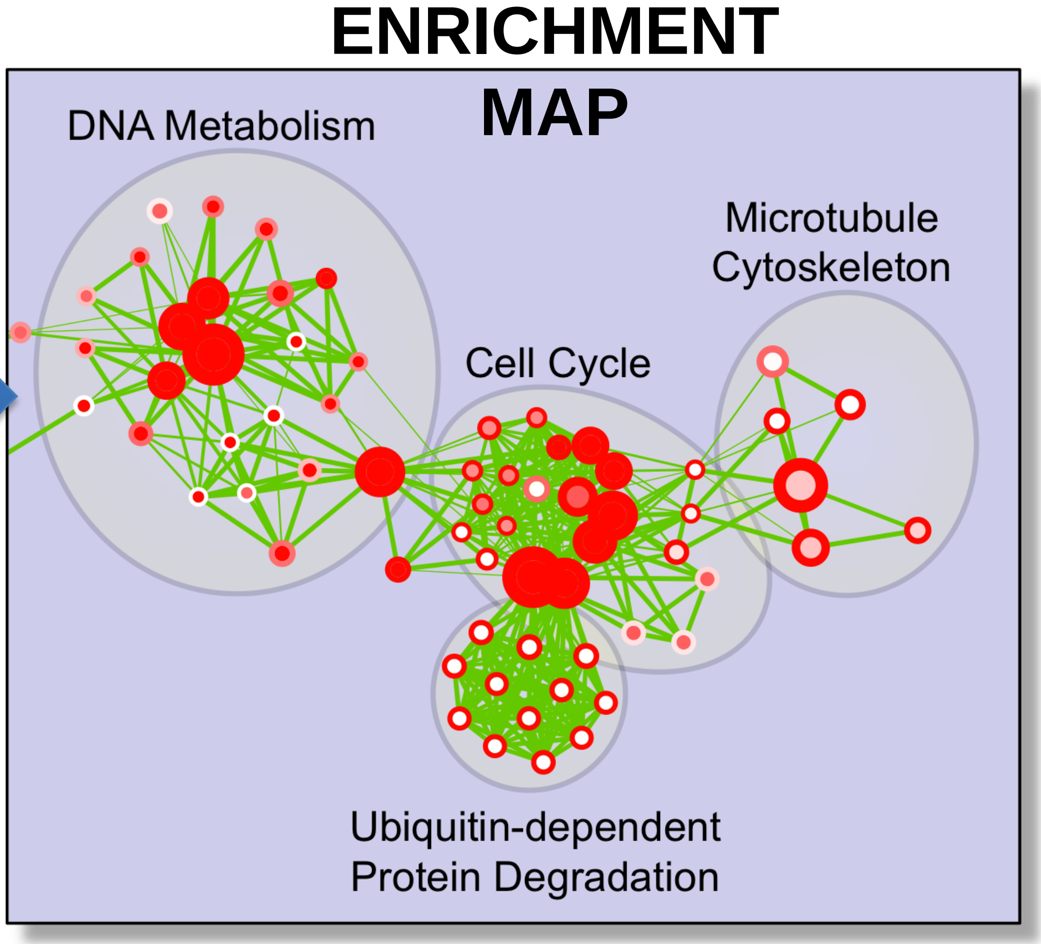
Gene.G
Gene.H
Gene.I

Apoptosis

Gene.D
Gene.E
Gene.F

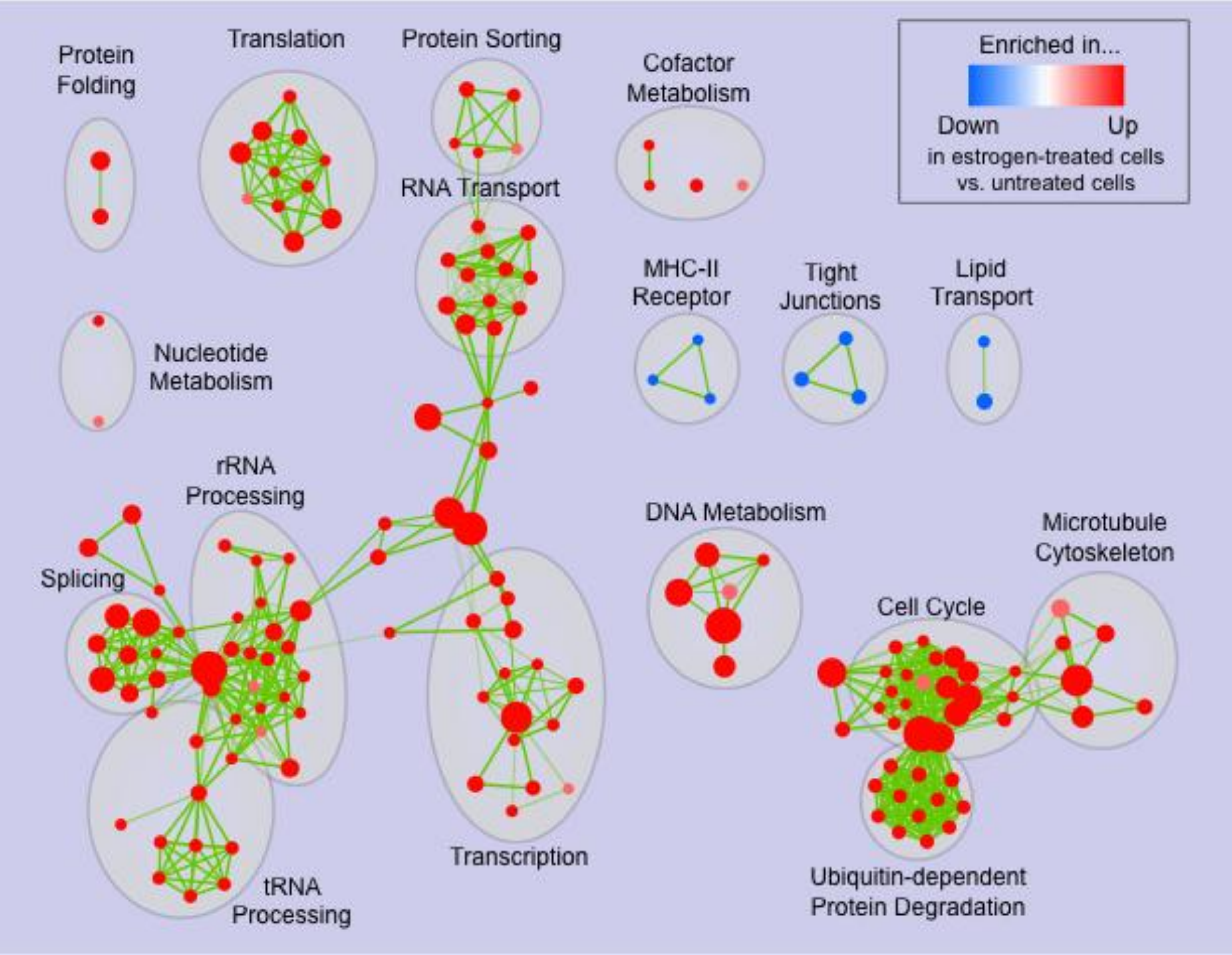
MAPK

Gene.L
Gene.M
Gene.N



- 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

ORA: topGO, clusterProfiler, RDAVIDWebService, ReactomePA, enrichR, GOseq, PathwaySplice

FCS: globaltest, gage, Camera, PADOG, SetRank

Others: GSVA, SPIA, PathNet, TcGSA, QuSAGE, DNEA

Ensembles: piano, EGSEA, ToPASeq…
And many more

Final remarks:

R
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- 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