Bioinformatics for Pathway Enrichment Analysis Part 1

Veronique Voisin and Ruth Isserlin

Computer Ontario Summer school 2025

Course outline

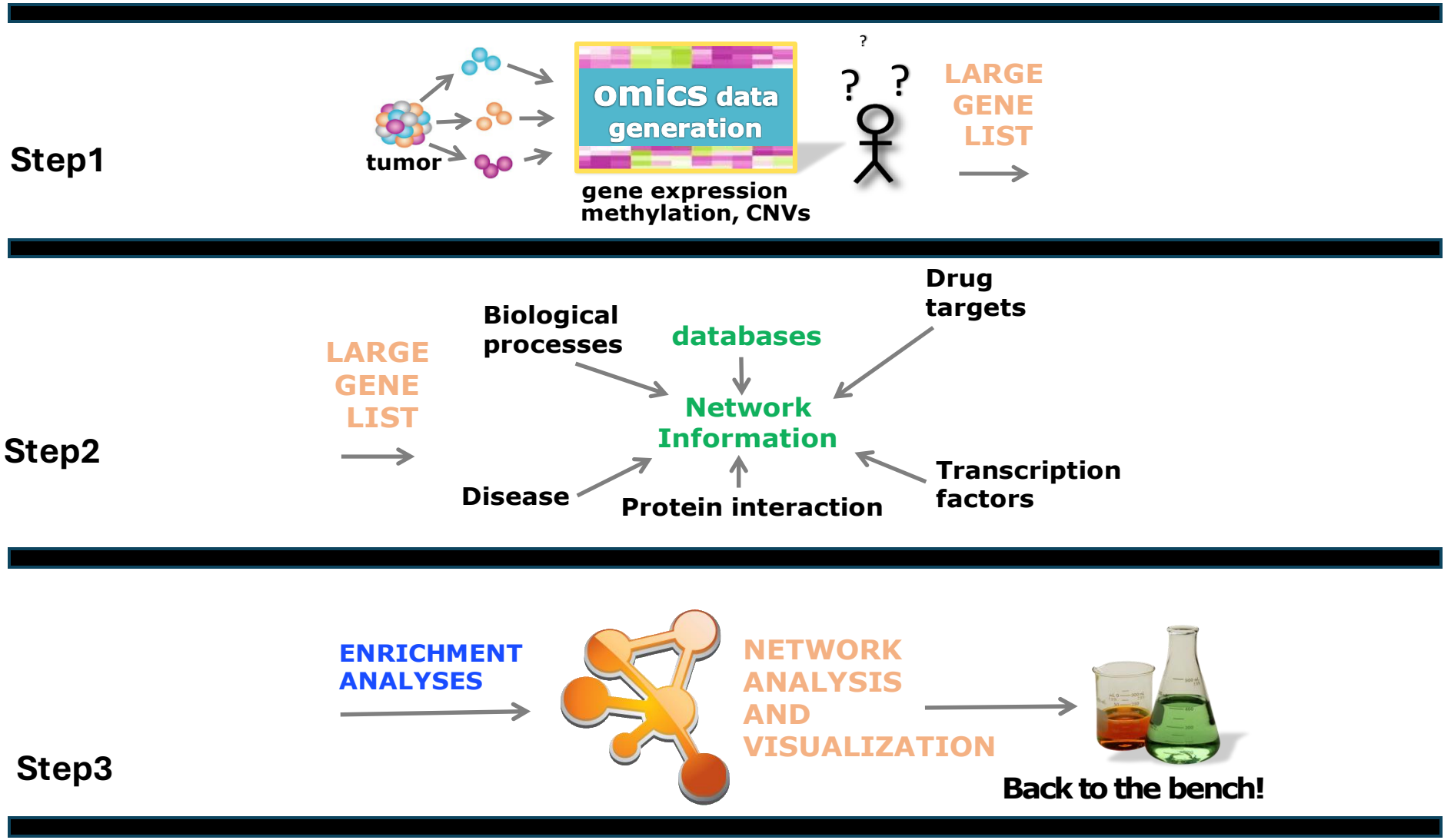
- General concept of pathway enrichment analysis using a defined gene list.
- What are the steps to perform pathway enrichment analysis ?
- Example : frequently mutated genes
- Practical lab: pathway enrichment analysis using R

Learning Objectives

- Be able to understand:

  - the main goal of pathway enrichment analysis
  - the general workflow of pathways enrichment analysis (what is the step1 (input data) and what is the output?)
  - the advantages of pathway enrichment analysis

General Workflow of pathway enrichment analysis



What Is a Gene List?

A gene list is a set of genes identified through an experiment or analysis. Shared features among genes in a list may suggest:

- **Involvement in the same biological system**
  (e.g., pathway, protein complex, or physical interactions)

- **Similar molecular function**
  (e.g., all encode protein kinases)
- **Expression in the same cell type or tissue**

- **Proximity on the genome**
  (e.g., due to linkage or copy number variations)

Where Do Gene Lists Come From?

Gene lists are generated through various types of **high-throughput experiments** and **computational analyses**:

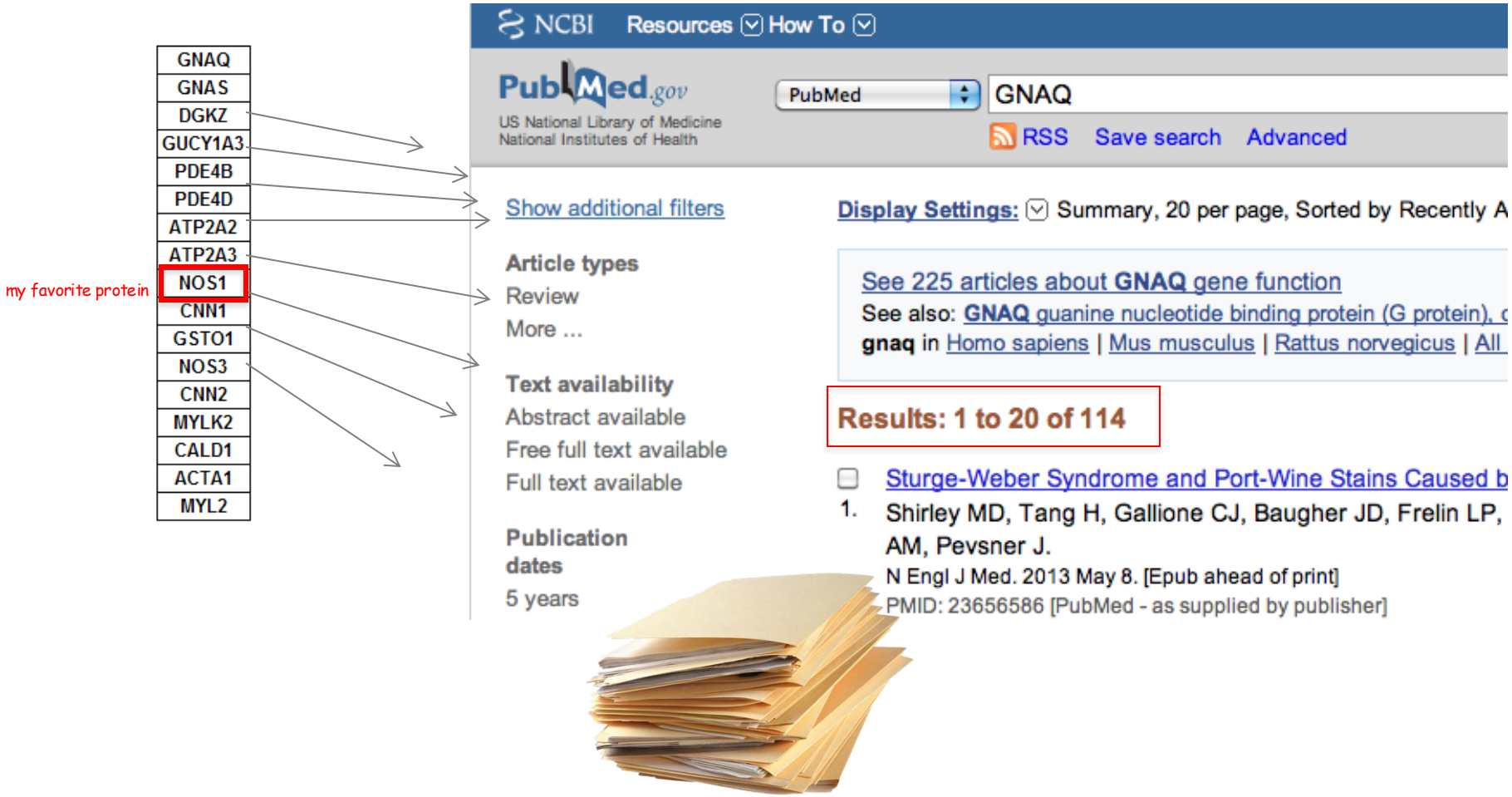
Molecular Profiling
- **Transcriptomics** (e.g., RNA-seq → gene expression changes)
- **Proteomics** (e.g., mass spectrometry → protein abundance)

Functional Genomics
- **Genetic screens** (e.g., CRISPR, RNAi, knockout libraries)
- **Perturbation studies** (e.g., drug treatment, environmental stress)

Genetic Variation Studies
- **Genome-wide association studies (GWAS) (**SNPs linked to disease or traits)
- **Structural variation (**Copy number variants (CNVs), insertions, deletions)

- **And more….**

Pathway enrichment analysis using bioinformatic tools save time compared to the traditional approach



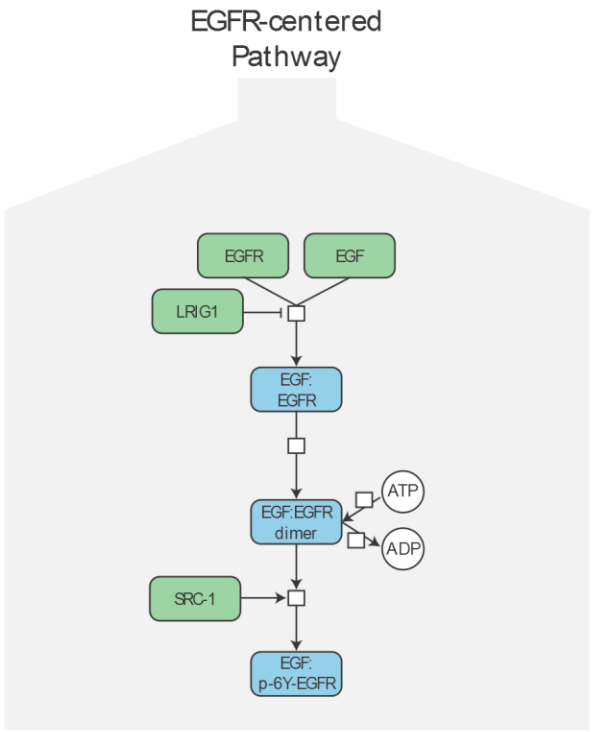
Benefits of Pathway Enrichment Analysis

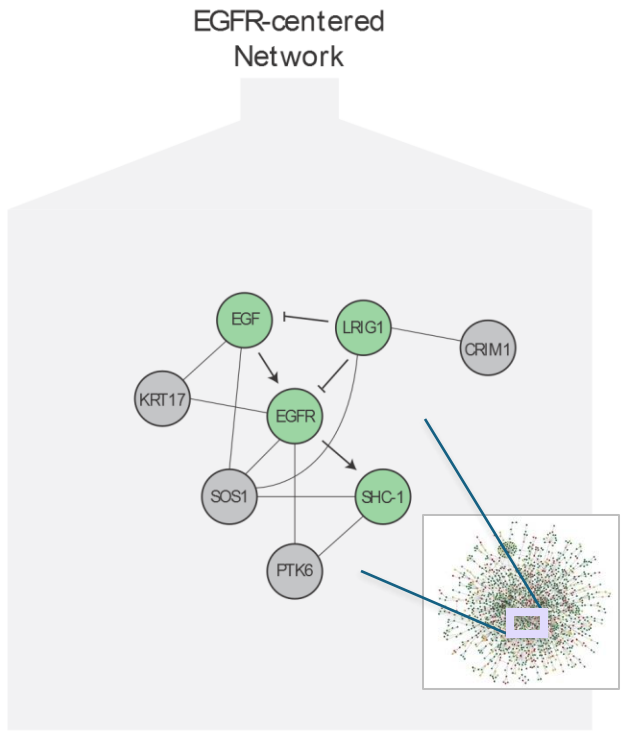
vs. analysis of individual transcripts, proteins, SNPs…

- Results are easier to interpret
  - Familiar concepts e.g. cell cycle
- Improves statistical power
  - Fewer tests, aggregates data from multiple genes into one pathway
- More reproducible
  - E.g. gene expression signatures
- Facilitates integration of multiple data types

- Identifies possible causal mechanisms
- Predicts new roles for genes

Pathways vs. Networks





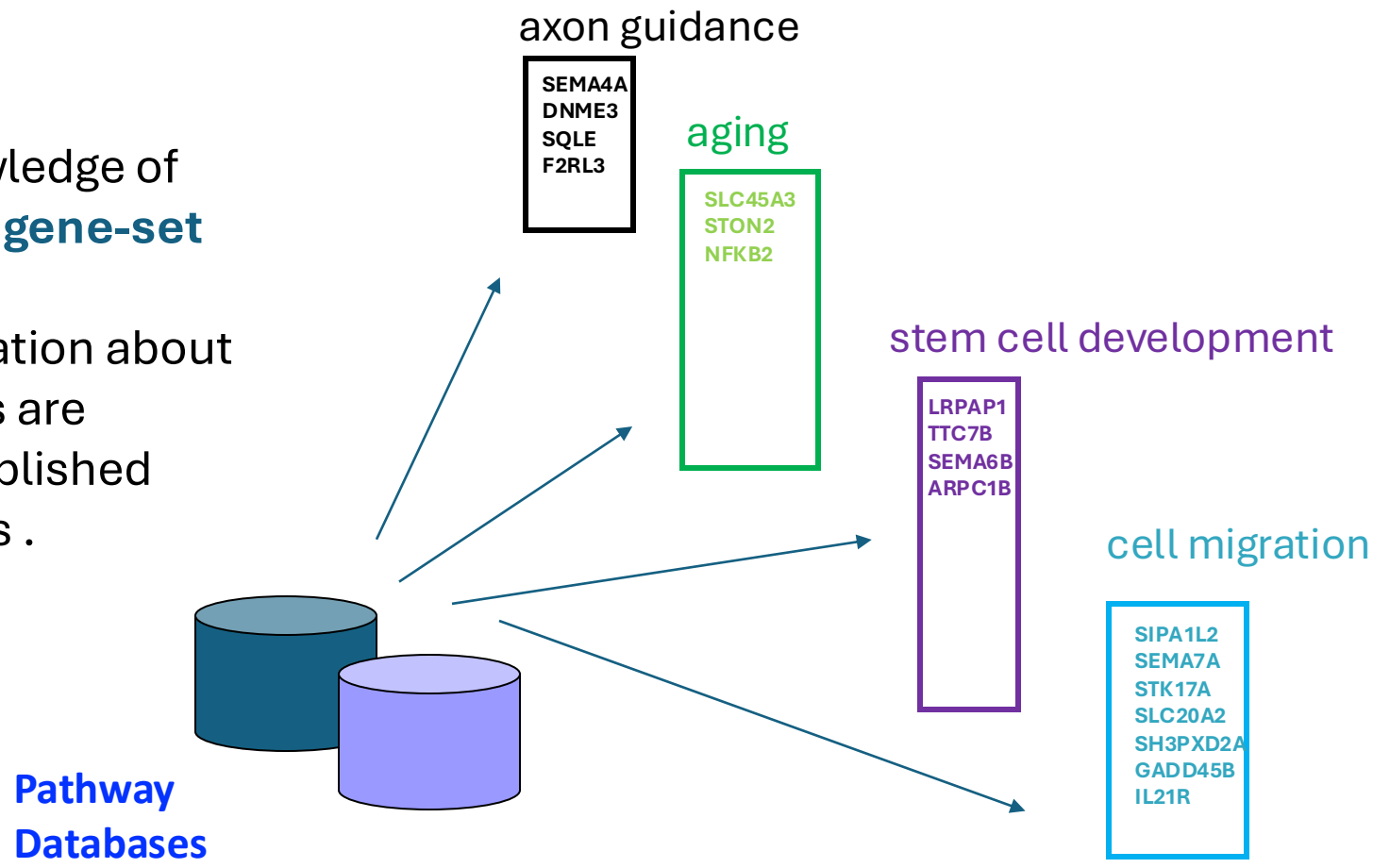
- Detailed, high-confidence consensus
- Biochemical reactions
- Small-scale, fewer genes
- Concentrated from decades of literature

- Simplified cellular logic, noisy
- Abstractions: directed, undirected
- Large-scale, genome-wide
- Constructed from *omics* data integration

Pathway: proteins that are known to work together in a defined biological process.
A network is a representation of a pathway, built using information stored in pathway database

Pathway database

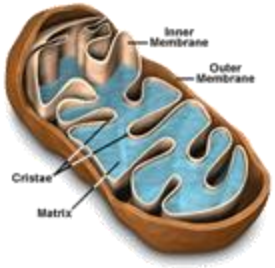
We use prior knowledge of pathway stored in **gene-set** format in pathway Database. Information about genes in pathways are collected from published papers by curators .

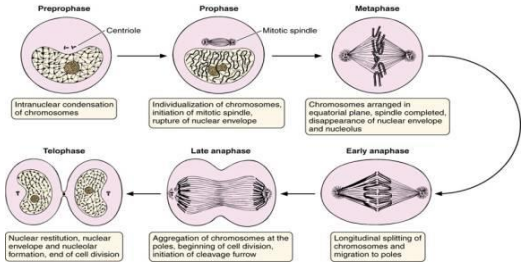


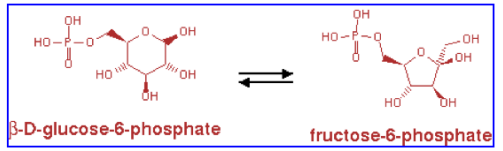
GENEONTOLOGY
Unifying Biology

What GO Covers?

- GO terms divided into three aspects:
  - cellular component
  - molecular function
  - biological process



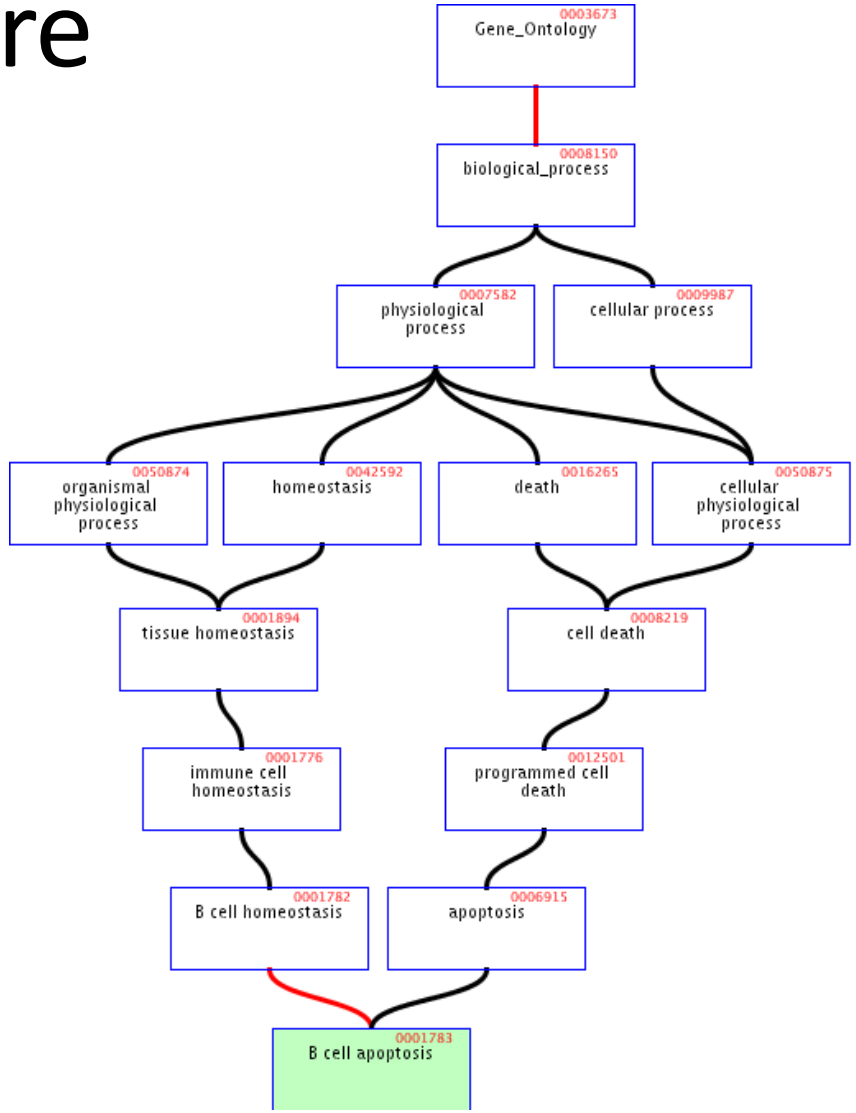




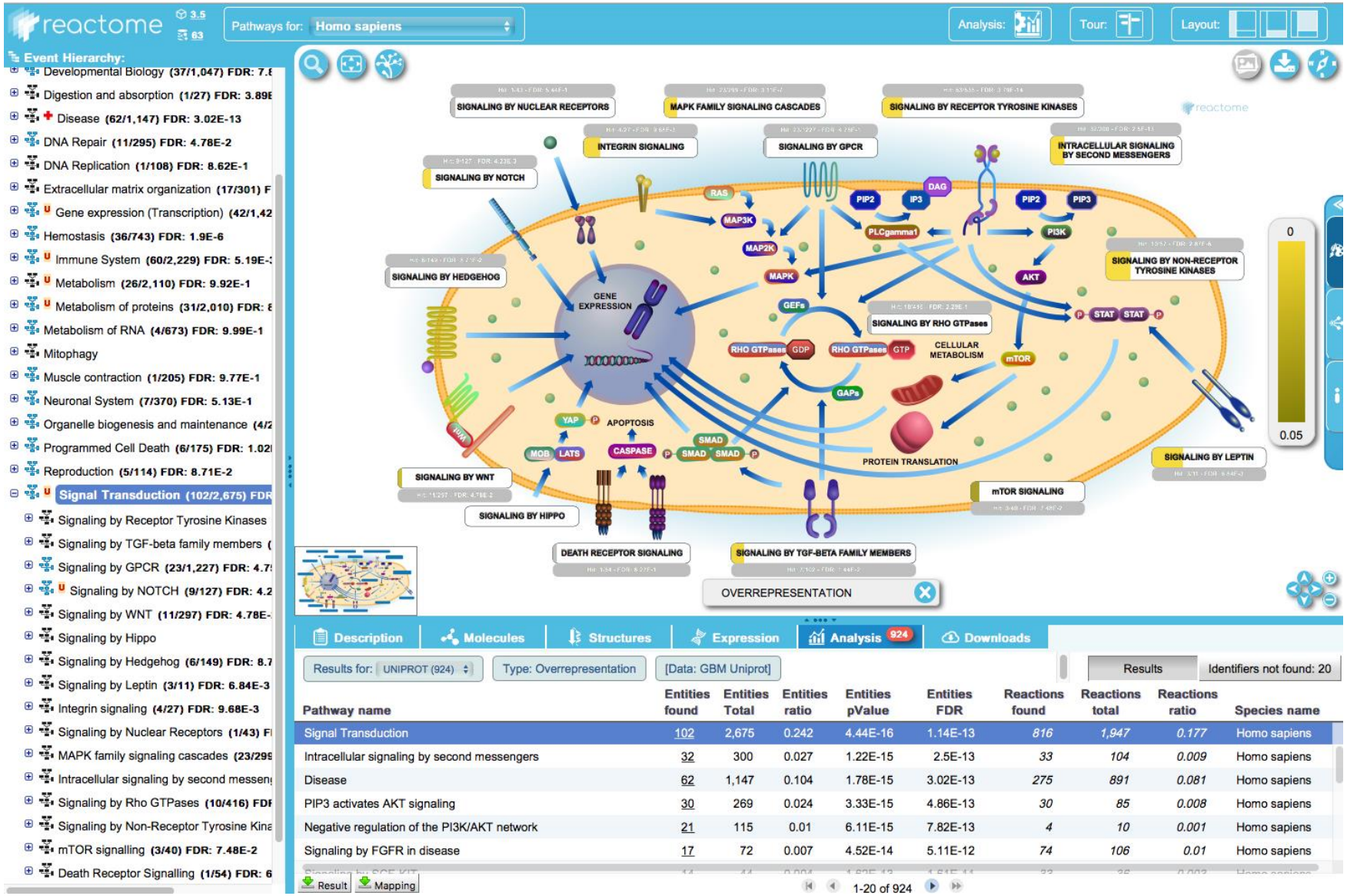
glucose-6-phosphate isomerase activity

GO Structure

- Terms are related within a hierarchy
  - is-a
  - part-of
- Describes multiple levels of detail of gene function
- Terms can have more than one parent or child



Another database: Reactome



What a Pathway File Looks Like : Displaying a Few Entries

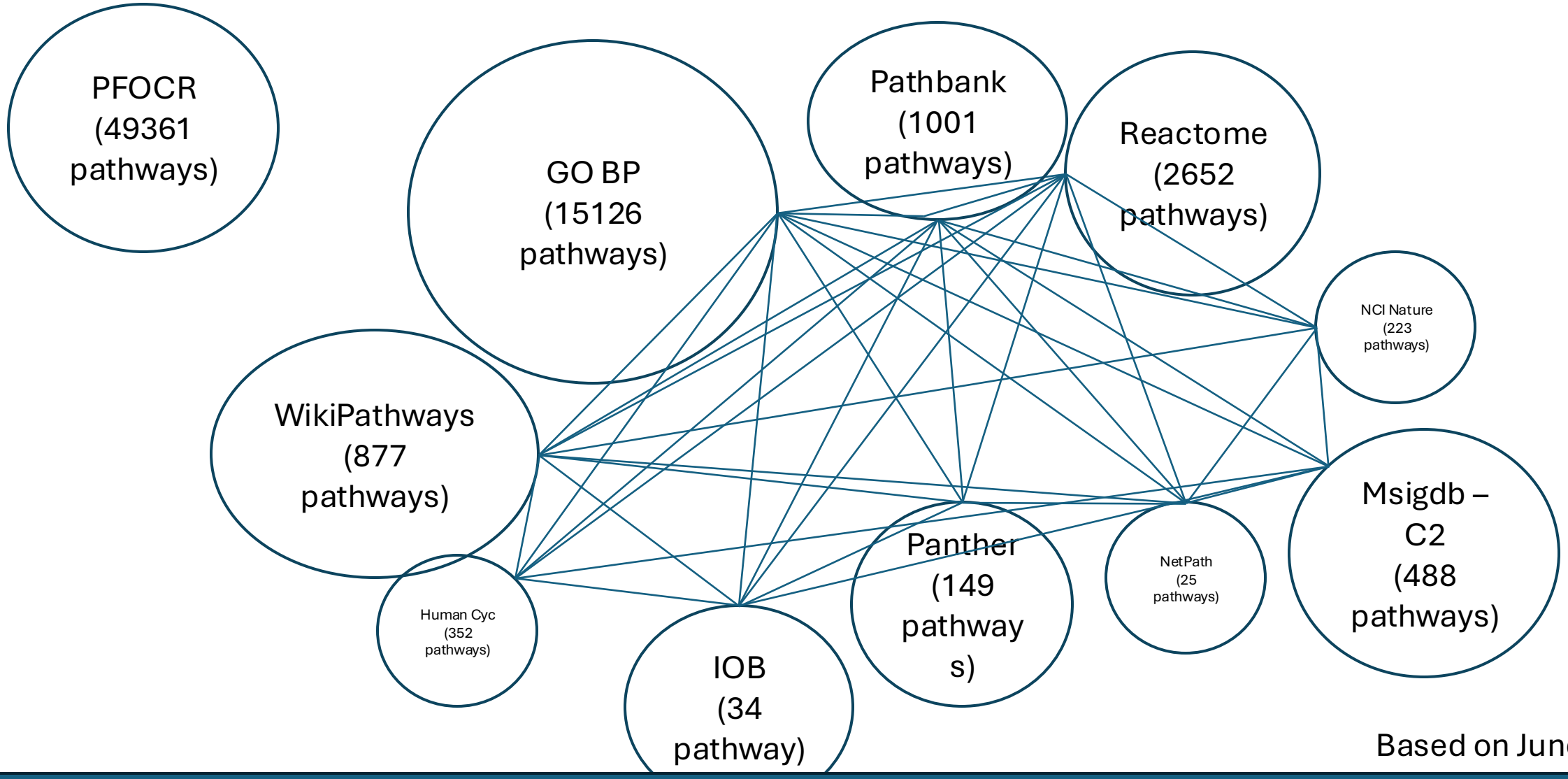
negative regulation of protein complex assembly	DACT1	ADD2	SOST	PFN1	ULK1	HEY2	MAPRE1	TUBB4A	TBCD
mesenchymal stem cell proliferation	SIX2								
interneuron axon guidance	LHX1	LHX9							
peptidyl-lysine oxidation	LOXL3	LOXL4	LOXL1	LOX	LOXL2				
negative regulation of skeletal muscle cell proliferation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

18000 pathways in file

Yellow: genes that match with my list

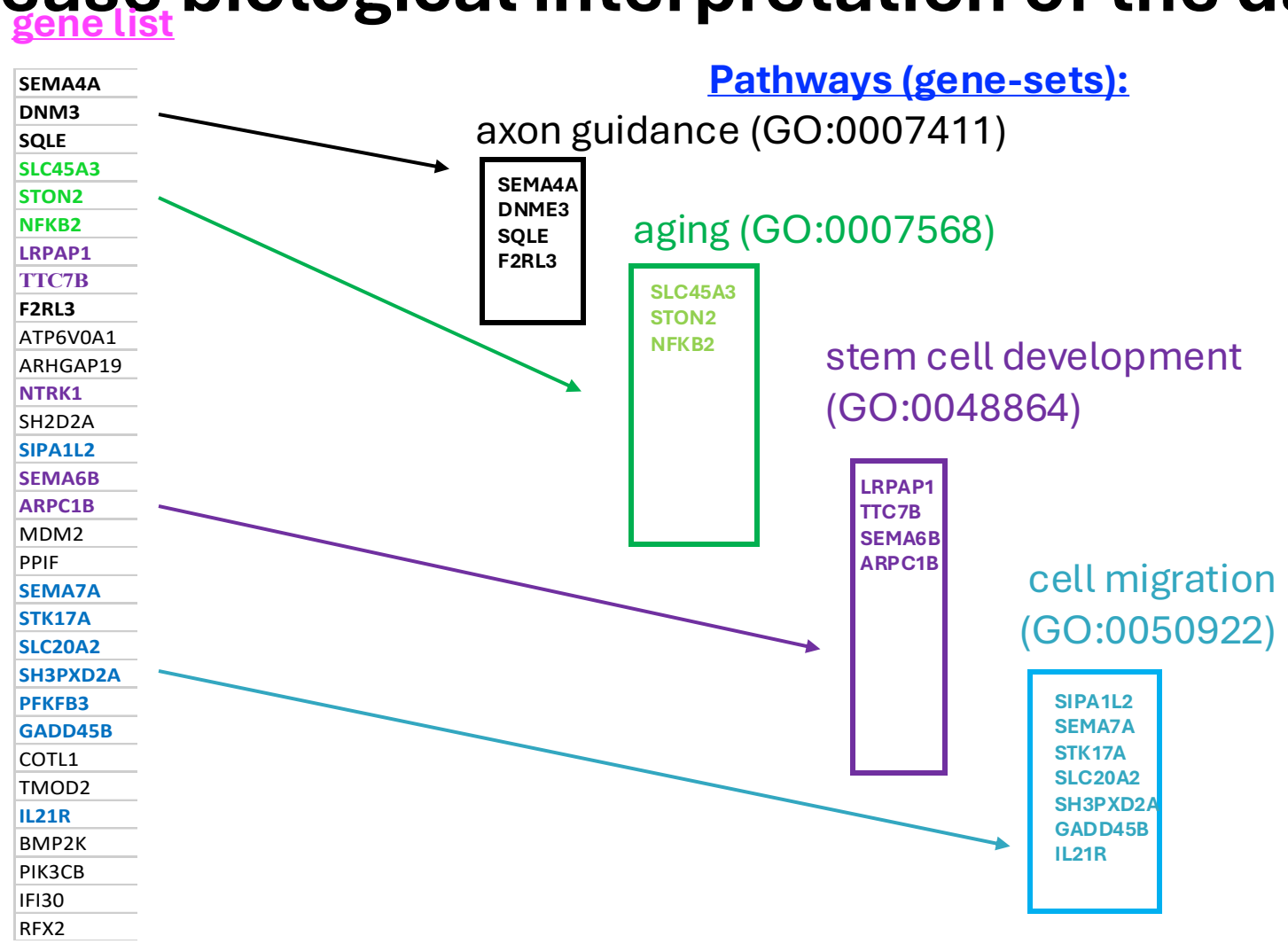
BaderLab EM_Genesets

http://download.baderlab.org/EM_Genesets/



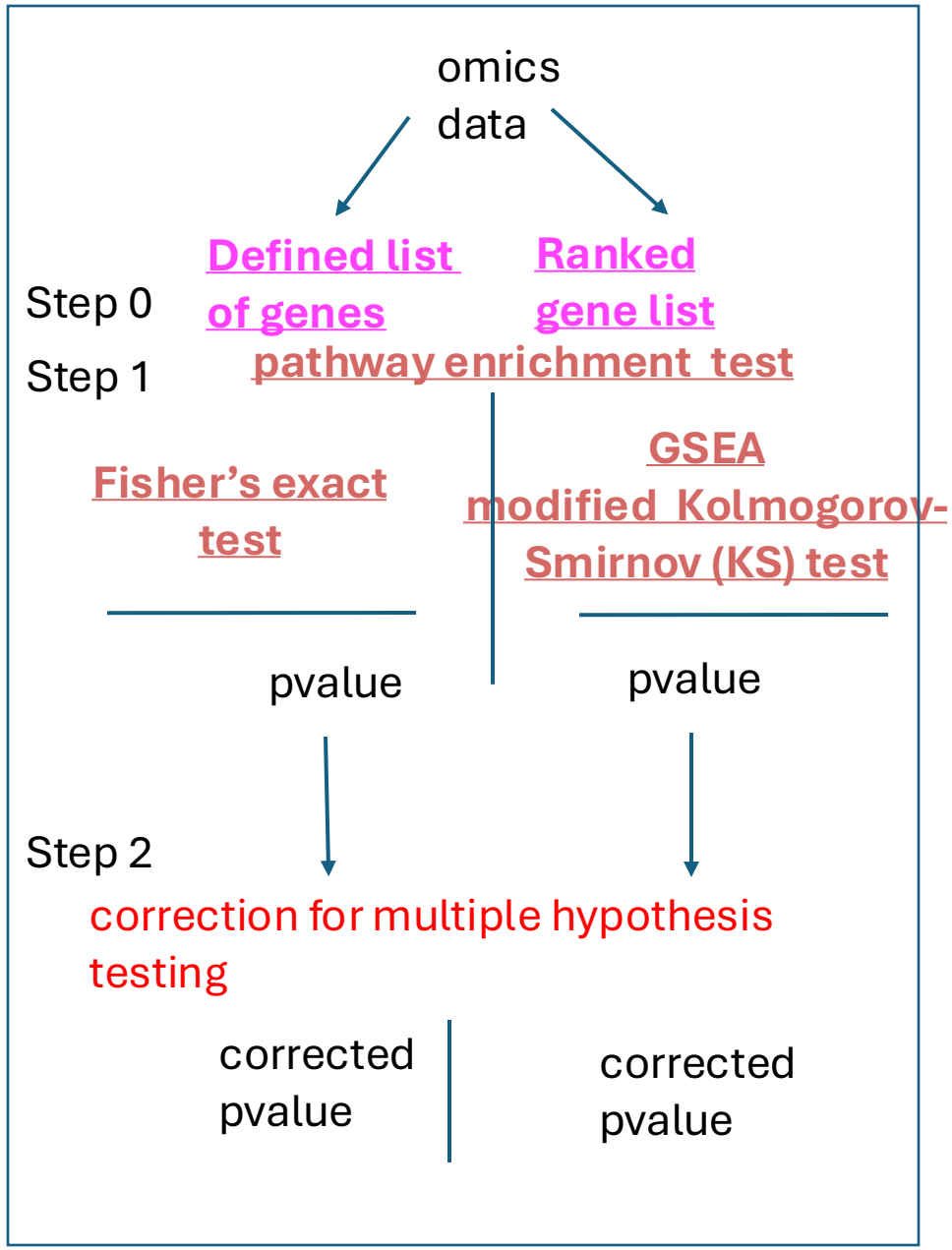
Based on June 2024 build

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



Two types of gene lists : defined or ranked gene list.

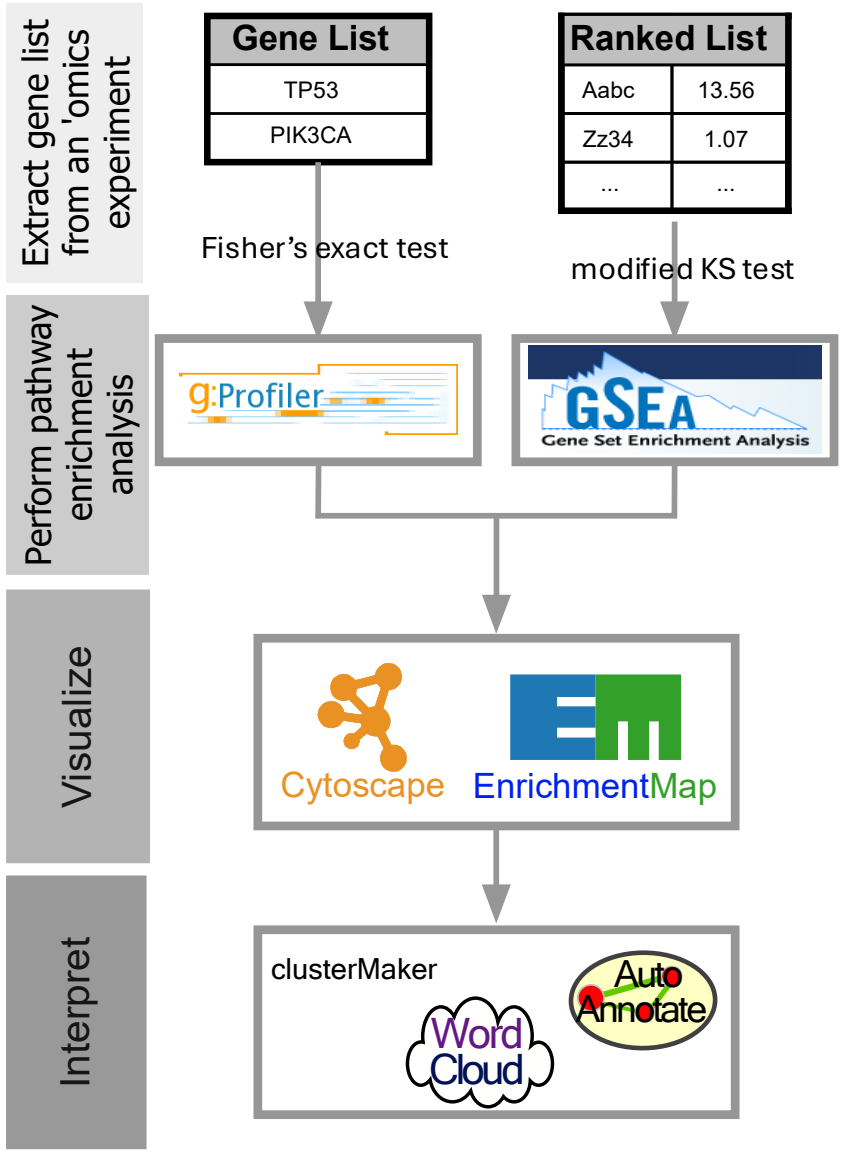
- Fisher's Exact Test, aka Hypergeometric Test
- GSEA for ranked lists.
- Multiple test corrections:
  - Bonferroni correction
  - False Discovery Rate computation using Benjamini-Hochberg procedure



Types of pathway enrichment analysis

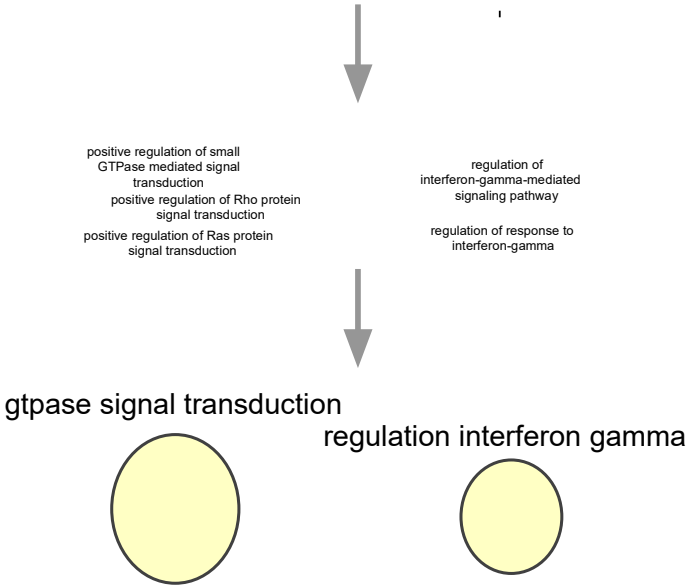
- Defined gene list (e.g. abundance change > 2-fold)
  - Answers the question: **Are any pathways (gene sets) surprisingly enriched in my protein list?**
  - Statistical test: **Fisher's Exact Test** (aka Hypergeometric test)
- Ranked gene list (e.g. by differential abundance)
  - Answers the question: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

Workflow



Outputs

Pathway	P-value	Q-value
POSITIVE REGULATION OF RAS PROTEIN SIGNAL TRANSDUCTION%GOBP%GO:0046579	0.00304414	0.0056384853
REGULATION OF INTERFERON-GAMMA-MEDIATED SIGNALING PATHWAY%GOBP%GO:0060334	0.0	0.0038799183
POSITIVE REGULATION OF RHO PROTEIN SIGNAL TRANSDUCTION%GOBP%GO:0035025	0.004622496	0.008516296



Pathway enrichment analysis using a defined gene list

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

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

gene list

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

• • •

FDR<0.05

My gene list

PFKB3 SH2D2A MDM2 RFX2 IFI30 SLC45A3 SIPA1L2 ARHGAP19 TMOD2 SEMA7A STON2 LRPAP1 SH3PXD2A SEMA6B COTL1 NFKB2 ATP6V0A1 GADD45B STK17A TTC7B ARPC1B SLC20A2 IL21R NTRK1 SLC20A2 PPIF BMP2K PIK3CB

overlap

Pathway (gene-set):

axon guidance (GO:0007411)

MYPN LAMC2 SEMA3G TUBB3 SLIT1 GAB2 LAMA1 SEMA4A RELN BDNF DOK4 DOK1 SEMA5B EFNB1 VASP PTPRA DNM3 SQLE BSG EPHB3 LAMA5 NRP2 F2RL3 SPTB BDNF IGSF9 EFNA2 SMO KIF5C TNR SEMA7A LGI1 GFRA3 LAMA2 ALCAM DOK4 PALLD PTPN11

13

Size of the original pathway **39**

Size of the gene list **41**

13/41 → ~1/3

13/39 → ~1/3

Size = number of genes

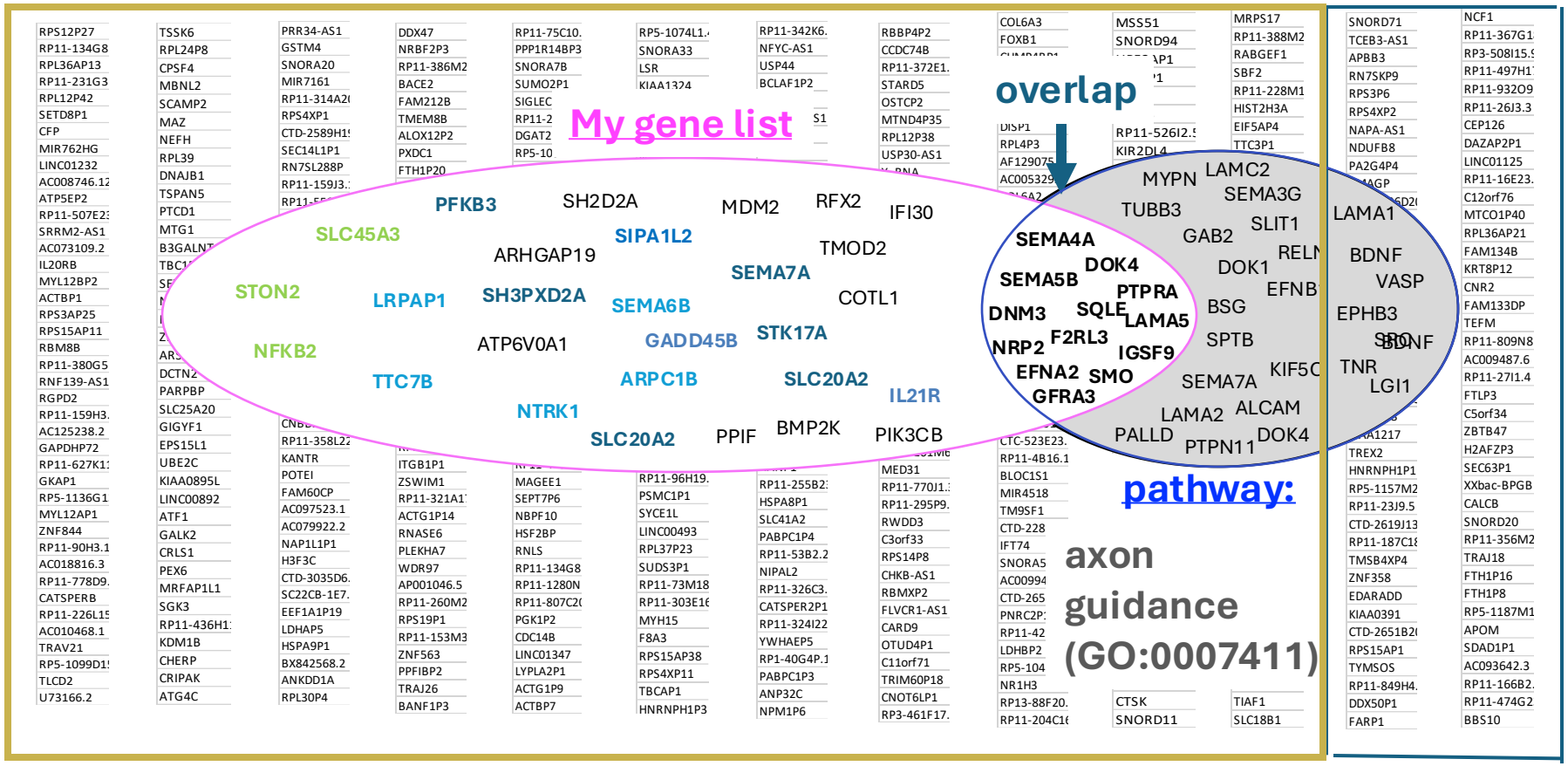
The background represents all the proteins that could have been captured in my omics experiment

gene list

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

genes measured in the experiment

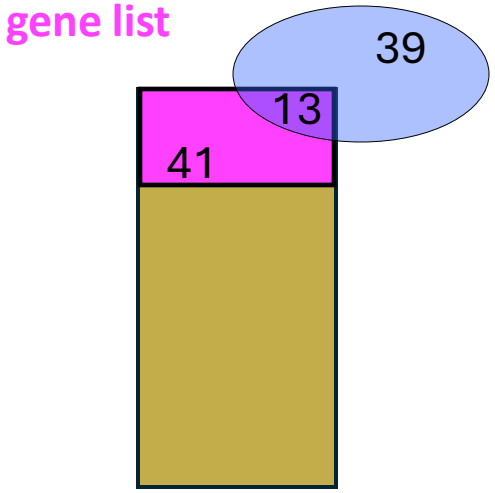
genes not measured



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

How many proteins could have been captured in your experiment? : about 6,000?

How do simple enrichment tests work?



OVERLAP BETWEEN GENE LIST AND GENE_SETS (PATHWAYS)

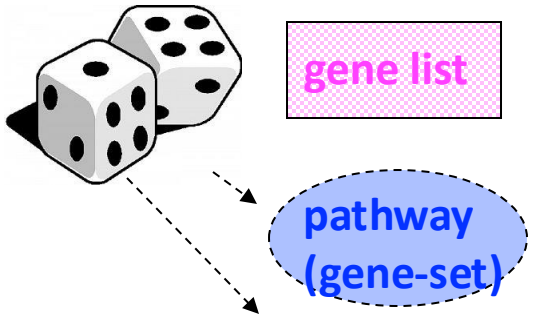
1

Enrichment Table

Gene-set	p-value
Spindle	0.0001
Apoptosis	0.025

2

Is this overlap larger than expected by chance?



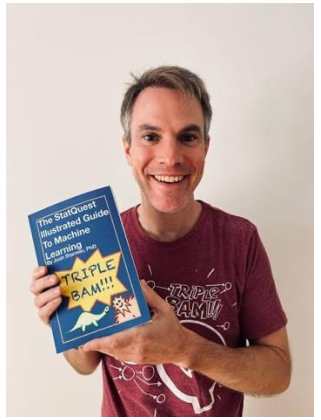
random sampling using background genes

Empirical pvalue

Background genes

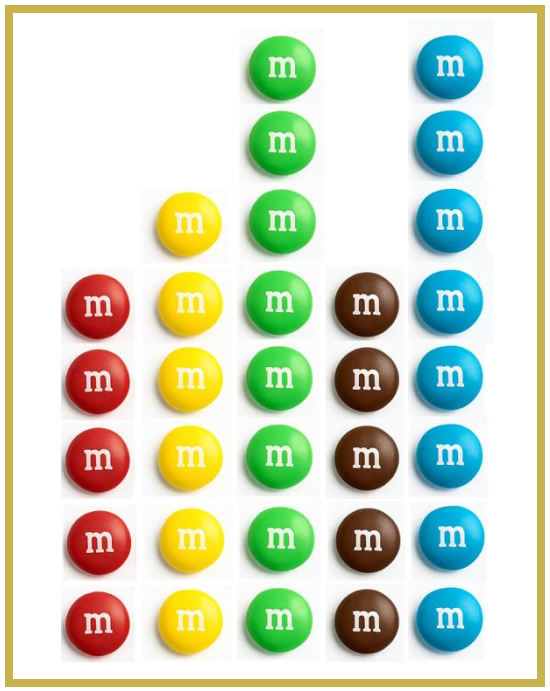
Do you need to learn more about Fisher's exact test?

StatQuest with Josh Starmer



1 bag of m&m's
(= my universe)

gene sets (pathways)(



Background

My "sample"= my gene list

On the left, all the m&m's in the bag sorted using the different colors. I take a random "sample" of m&m's from the bag, and I would like to determine if my sample is special.

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

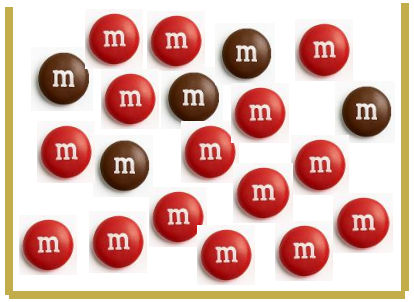
The Fisher's exact test

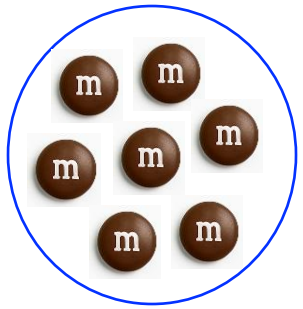
a.k.a., hypergeometric test

"my sample"

Gene list

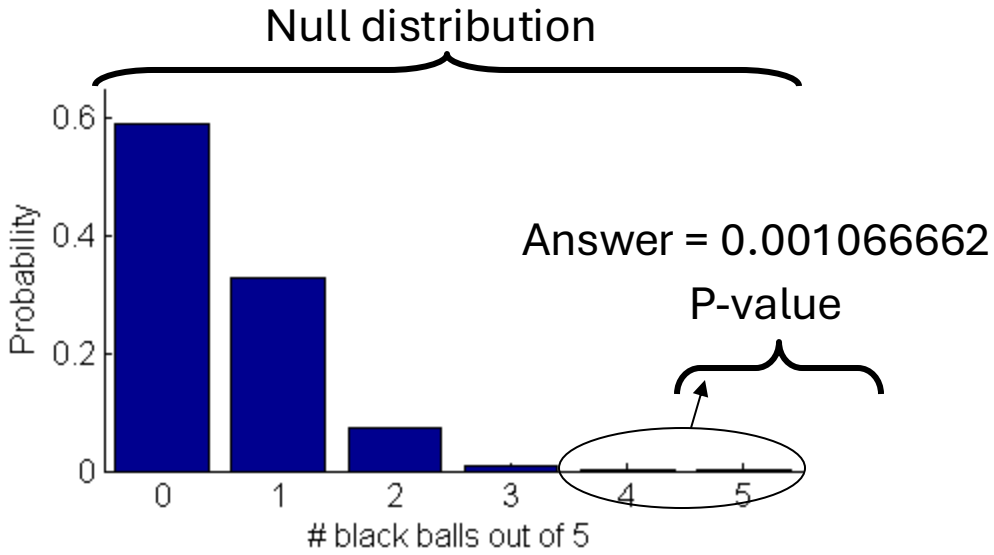
RRP6
MRD1
RRP7
RRP43
RRP42

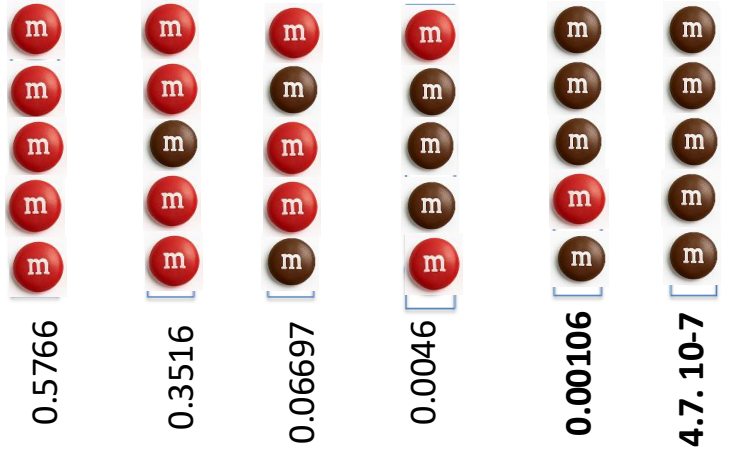




Background population:
500 black m&m's,
4500 red m&m's

1 gene-set/pathway

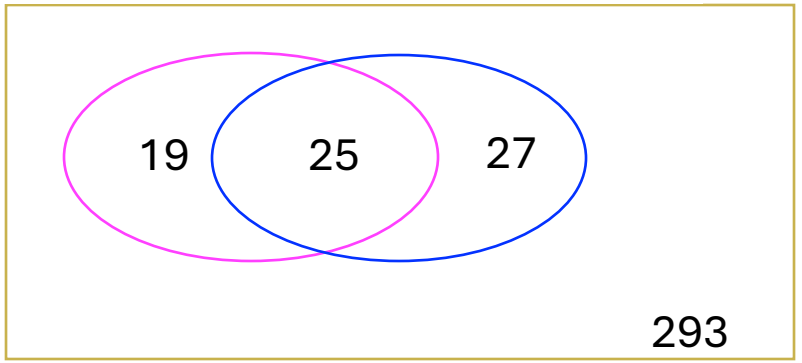




g:Profiler

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

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



2x2 contingency table

	In protein list	Not in protein list
In pathway	25	27
Not in pathway	19	266

Fisher's exact test

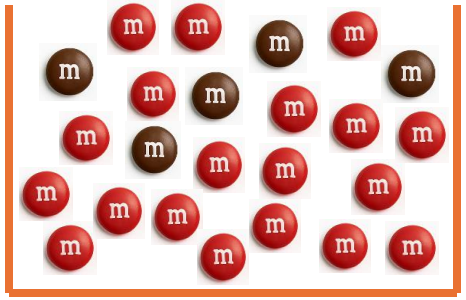
We are testing many pathways at the same time

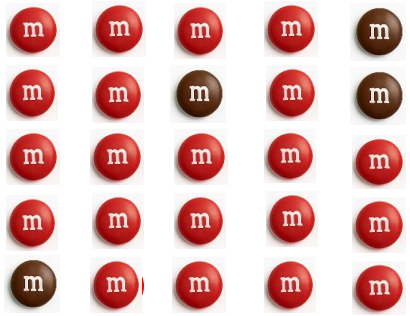
➡ We need to correct for multiple hypothesis testing

Correction for Multiple Hypothesis Testing

How to win the p-value lottery

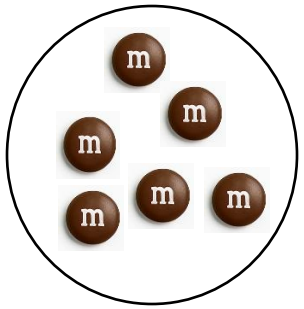
Random draws





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

… 7,834 draws later …



Background population:
500 black genes,
4500 red genes

1 gene-set
(apoptosis)

Correction for Multiple Hypothesis Testing

1. Simple P-value correction: Bonferroni

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

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

2. False discovery rate (FDR)

FDR is *the expected **proportion** of the observed enrichments due to random chance.*

- Typically, FDR corrections are calculated using the Benjamini-Hochberg procedure.
- FDR threshold is often called the "q-value" or "adjusted pvalue"

NAME	SIZE	ES	NES	NOM p-val	FDR q-val
COFACTOR CATABOLIC PROCESS%GOBP%GO:0051187	25	0.7750398	2.368539	0	5.02E-04
ANTIBIOTIC CATABOLIC PROCESS%GOBP%GO:0017001	21	0.78641075	2.2833898	0	0.00219077
HYDROGEN PEROXIDE METABOLIC PROCESS%GOBP%GO:0042743	16	0.8221468	2.2569728	0	0.00275445
PEPTIDE CHAIN ELONGATION%REACTOME%R-HSA-156902.2	75	0.5671107	2.19405	0	0.00696158
RESPONSE OF EIF2AK4 (GCN2) TO AMINO ACID DEFICIENCY%REACTOME%R-HSA-96	82	0.56710726	2.1913495	0	0.00576
EUKARYOTIC TRANSLATION ELONGATION%REACTOME%R-HSA-156842.2	78	0.56959534	2.1875012	0	0.00536845
VIRAL MRNA TRANSLATION%REACTOME%R-HSA-192823.3	76	0.56925994	2.180141	0	0.00501747
HALLMARK_SPERMATOGENESIS%MSIGDB_C2%HALLMARK_SPERMATOGENESIS	45	0.61634994	2.1523428	0	0.00671322
L13A-MEDIATED TRANSLATIONAL SILENCING OF CERULOPLASMIN EXPRESSION%RE	95	0.5357858	2.1477811	0	0.00651045
EUKARYOTIC TRANSLATION INITIATION%REACTOME DATABASE ID RELEASE 72%72(	103	0.52849996	2.1412485	0	0.00688575
EUKARYOTIC TRANSLATION TERMINATION%REACTOME%R-HSA-72764.4	79	0.56097984	2.1397927	0	0.00634908
CAP-DEPENDENT TRANSLATION INITIATION%REACTOME DATABASE ID RELEASE 72%	103	0.52849996	2.1329718	0	0.00658981
GTP HYDROLYSIS AND JOINING OF THE 60S RIBOSOMAL SUBUNIT%REACTOME DAT	97	0.535648	2.132473	0	0.0060829

Extract the pathways that are significant at FDR 0.05 or less

EXAMPLE WITH A DEFINED GENE LIST

Frequently mutated genes

Article | Open access | Published: 16 October 2013

Mutational landscape and significance across 12 major cancer types

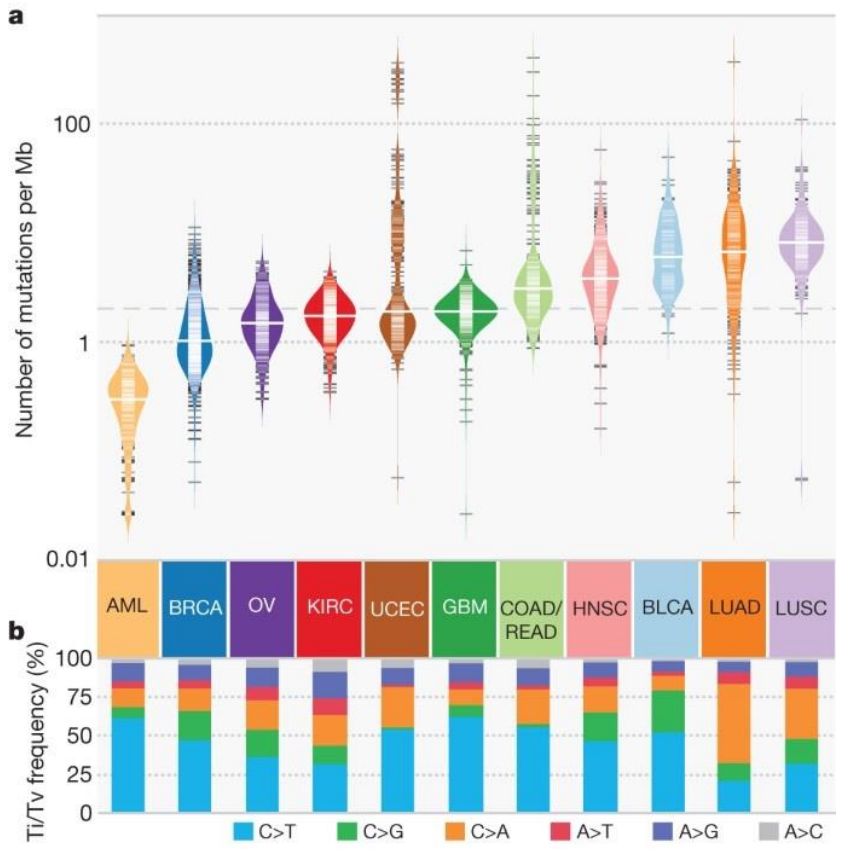
Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D. M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael & Li Ding

Nature **502**, 333–339 (2013) | Cite this article

201k Accesses | **228** Altmetric | Metrics

Data used for practical lab:

Dataset: Mutational landscape and significance across 12 major cancer types



https://www.nature.com/articles/nature12634 (2013)

Exome sequencing
Tumor samples and
matched control tissues

↓

Detection of points mutations and small insertions/deletions: somatic variant calls in each cancer type and in each tumor

↓

Calculation of mutation frequency: genes mutated in at least 5% of tumors were selected

↓

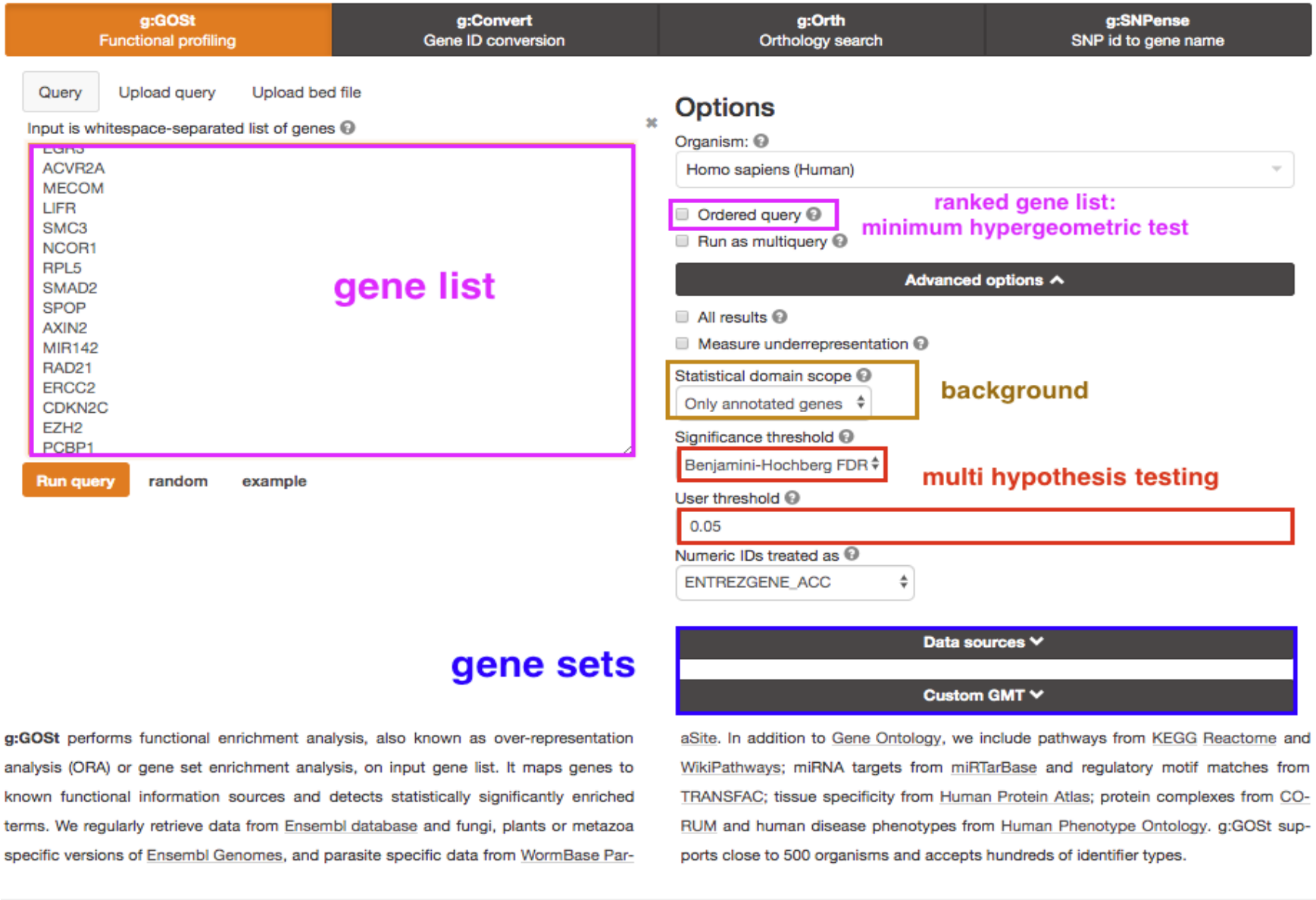
Genes positively correlated with number of mutation per sample

↓

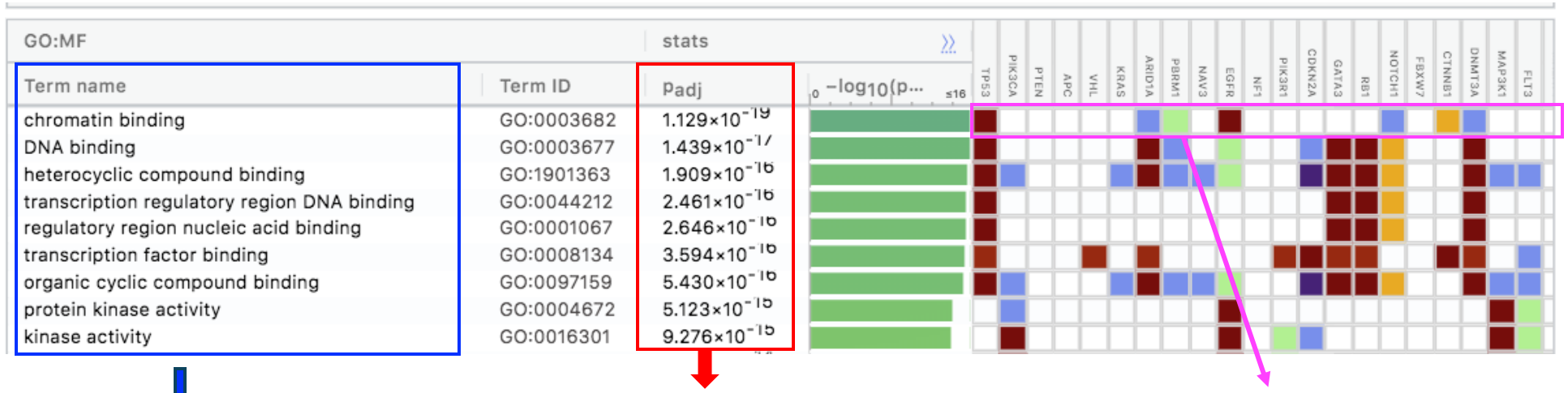
127 'significantly mutated genes'

gene list

g:Profiler


g:GOSt
Functional profiling
g:Convert
Gene ID conversion
g:Orth
Orthology search
g:SNPense
SNP id to gene name
Query
Upload query
Upload bed file
Input is whitespace-separated list of genes
ACVR2A
MECOM
LIFR
SMC3
NCOR1
RPL5
SMAD2
SPOP
AXIN2
MIR142
RAD21
ERCC2
CDKN2C
EZH2
PCBP1
gene list
Run query
random
example
Options
Organism:
Homo sapiens (Human)
Ordered query
Run as multiquery
ranked gene list:
minimum hypergeometric test
Advanced options
All results
Measure underrepresentation
Statistical domain scope
Only annotated genes
background
Significance threshold
Benjamini-Hochberg FDR
multi hypothesis testing
User threshold
0.05
Numeric IDs treated as
ENTREZGENE_ACC
gene sets
Data sources
Custom GMT
g:GOSt performs functional enrichment analysis, also known as over-representation analysis (ORA) or gene set enrichment analysis, on input gene list. It maps genes to known functional information sources and detects statistically significantly enriched terms. We regularly retrieve data from Ensembl database and fungi, plants or metazoa specific versions of Ensembl Genomes, and parasite specific data from WormBase ParaSite. In addition to Gene Ontology, we include pathways from KEGG Reactome and WikiPathways; miRNA targets from miRTarBase and regulatory motif matches from TRANSFAC; tissue specificity from Human Protein Atlas; protein complexes from CORUM and human disease phenotypes from Human Phenotype Ontology. g:GOSt supports close to 500 organisms and accepts hundreds of identifier types.

Explore results



each row is a gene-set (pathway)

Result of Fisher's exact test + multiple hypothesis correction: gene-sets (pathways) are significantly enriched at FDR <0.05 (scientific notation: 5×10^{-2})

colored boxes: genes in our gene list that overlap with the tested gene-set

Note: observe that same genes are included in several enriched gene-sets (pathways).

During the practical lab:
We will run g:Profiler from R

Summary of the Steps

1) Get our gene list

- Identification of genes with mutations present in at least 5% of the tumor samples
- Format gene list using official gene symbol

2) Perform pathway enrichment analysis (g:Profiler) or other equivalent tools

- the first step is to calculate the overlap size between each tested pathway and our list of proteins
- the second step is to estimate the probability (p value) that the enrichment happened by chance only using a Fisher's exact test
- the pvalue is corrected for multiple hypothesis testing (adjusted p value)
- the output is a list of pathways that are enriched in our list of proteins; we select the pathways enriched with a corrected pvalue less than 0.05 .

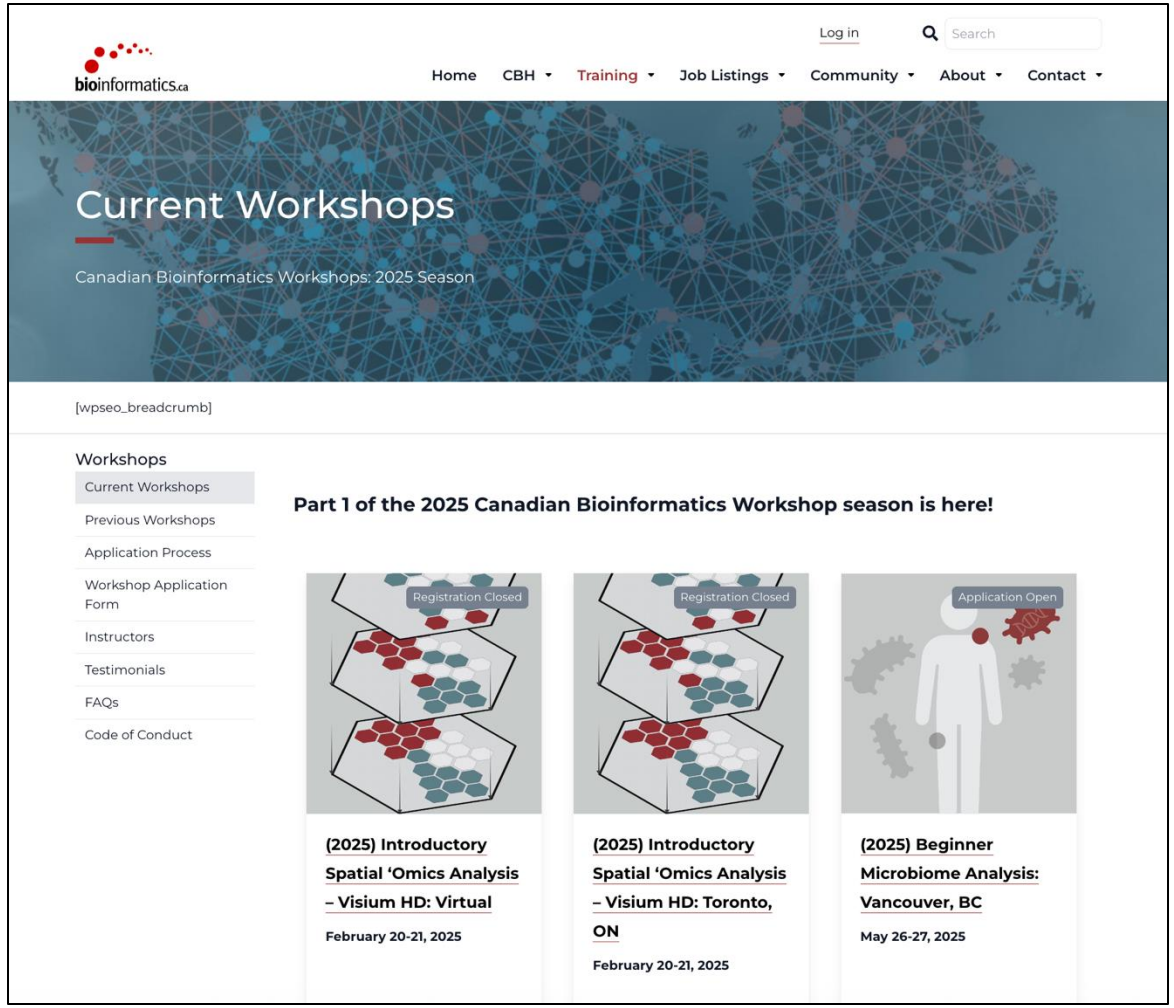
3) Visualized the results as a network (optional, see part2)

- The output table contains a lot of pathways that are similar and share some proteins in common; the table might be difficult to read or interpret.
- We can visualize the table as a network which will group the similar pathways together and enables us to identify the main groups of biological functions enriched in our gene list.

This module is taken from
CBW Pathway and Network Analysis Workshop 2024

Original source of the slides: Quaid Morris/Gary Bader/Lincoln Stein/Juri Reimand

2025 Canadian Bioinformatics Workshop
Please visit the website to see workshops presently offered



bioinformatics.ca

https://bioinformatics.ca/workshops/current-workshops/

Time for practical lab