Introduction to Bioinformatics

Topic 13
Pathway Enrichment Analysis

October 31st, 2023

Gregory R. Grant

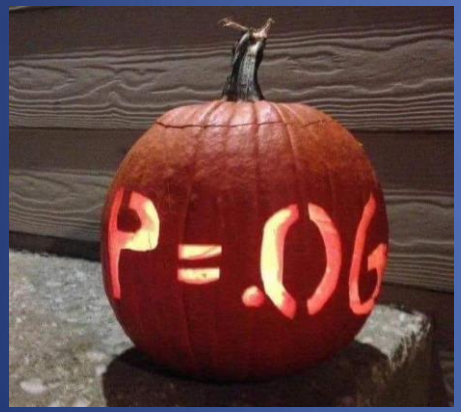
Genetics Department

ggrant@pennmedicine.upenn.edu

ITMAT Bioinformatics Laboratory

University of Pennsylvania

Professor
Gregory R. Grant



Happy Halloween

Teaching Assistants
Chetan Vadali
Jianing Yang

What comes after *q*-values?

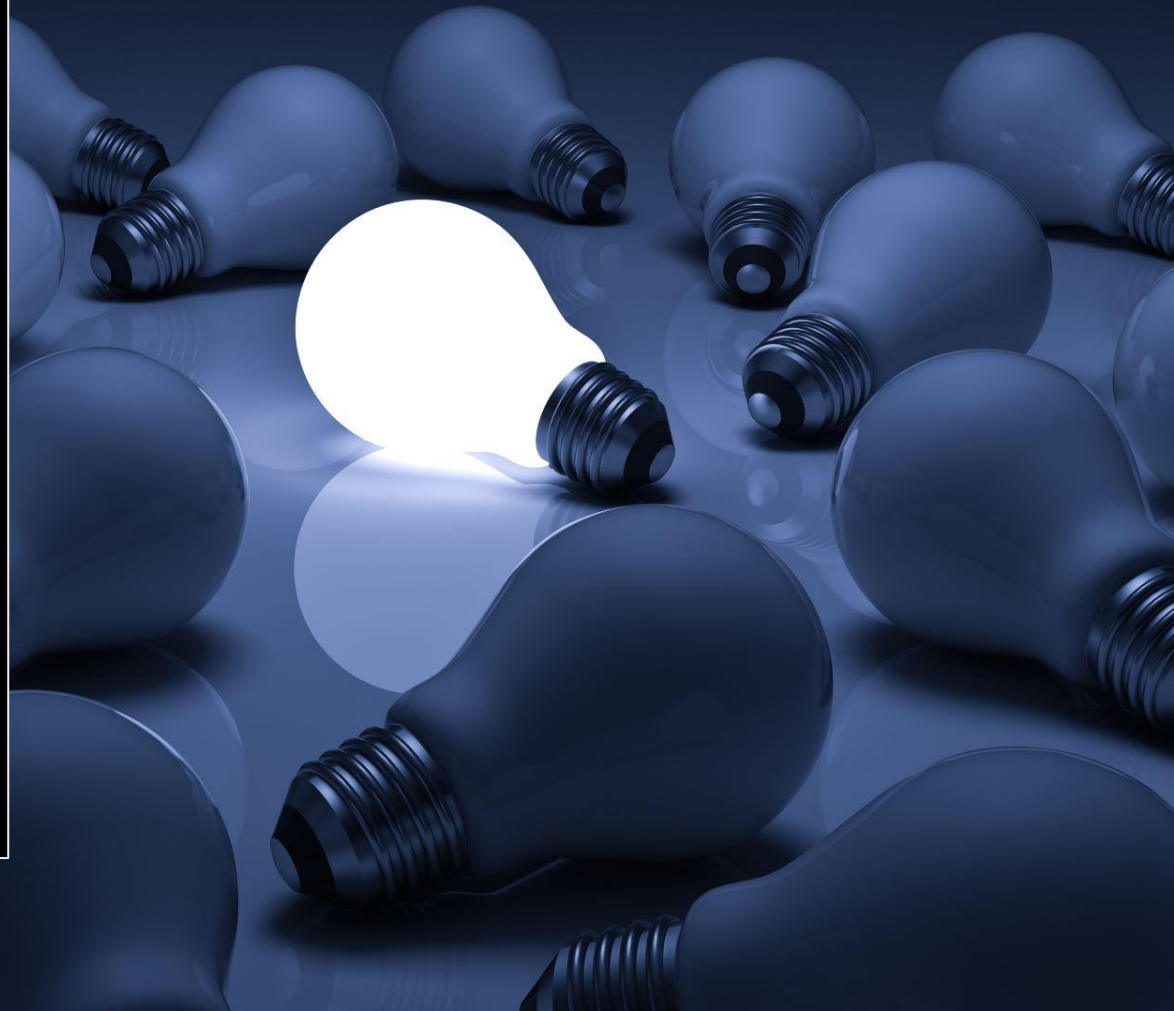
- What is done after *q*-values have been computed in an RNA-Seq DE analysis?
  - Case 1: All *q*-values equal 1, or are close to it, then we cannot reasonably conclude any genes are DE.
    - In spite of how low some *p*-values might seem.
  - Case 2: Only a few *q*-values are small, like 10 or 20, and all other *q*-values are 1 or close to it.
  - Case 3: Many genes, like hundreds or thousands of *q*-values are small, so hundreds or thousands of genes are DE.

Case 1

- No genes appear DE
  - Do not conclude no genes are DE.
  - We just failed to reject the null hypothesis.
  - Could simply be a matter of power.
- Were there enough replicates?
- Could power be increased by revisiting normalization, searching for batch effects, running a different DE statistical test (or app)?

Case 2

- There are just a few genes
  - Few enough to be investigated individually.
- Look up literature on them.
  - Do perturbation studies.
    - Knock them down, or out.
  - Look for genes they interact with.
  - Etc.



Case 3

- There are hundreds or thousands of DE genes.
  - This is very often the case.
  - It's not practical to investigate them individually.
  - Might drill down on some interesting looking ones.
    - It can take months if not years to investigate one gene.
    - This won't get you very far down a list of hundreds unless you get lucky and chase exactly the right gene.
- Instead, when there are hundreds, they typically fall into categories.
  - It's not typically genes that are DE between conditions, it's pathways.

Categories of Genes

- A category is a set of genes with some meaningful relation.

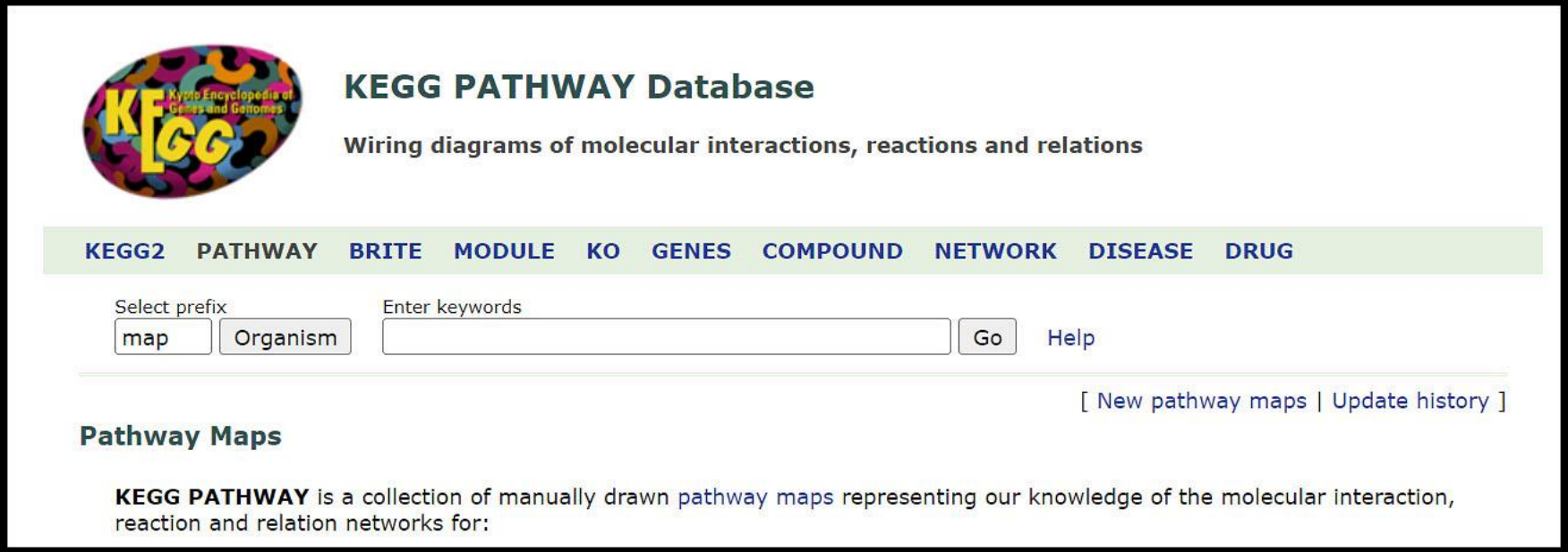
- There are many types of categories.
  - They can be functional (*e.g.,* clock genes), structural (*e.g.,* ribosomal genes), involved in the same processes (*e.g.,* cell cycle genes), etc.
  - Each category has multiple genes, and each gene will belong to multiple categories.
    - Genes/Categories is a many-to-many relationship.

- There are several working groups that have undertaken the task of categorizing genes.

- GO for example.


GENEONTOLOGY
Unifying Biology
About
Ontology
Annotations
Downloads
Help
ALLIANCE
Current release 2022-10-07: 43,329 GO terms | 7,694,564 annotations
1,503,740 gene products | 5,257 species (see statistics)
THE GENE ONTOLOGY RESOURCE
The mission of the GO Consortium is to develop a comprehensive, computational model of biological systems, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.
The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.
Search GO term or Gene Product in AmiGO ...
Any
Ontology
Gene Product
GO Enrichment Analysis
Powered by PANTHER
Your gene IDs here...
biological process
Homo sapien
Examples
Launch
Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs

GO


KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions and relations
KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND NETWORK DISEASE DRUG
Select prefix
map
Organism
Enter keywords
Go
Help
[New pathway maps | Update history]
Pathway Maps
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:

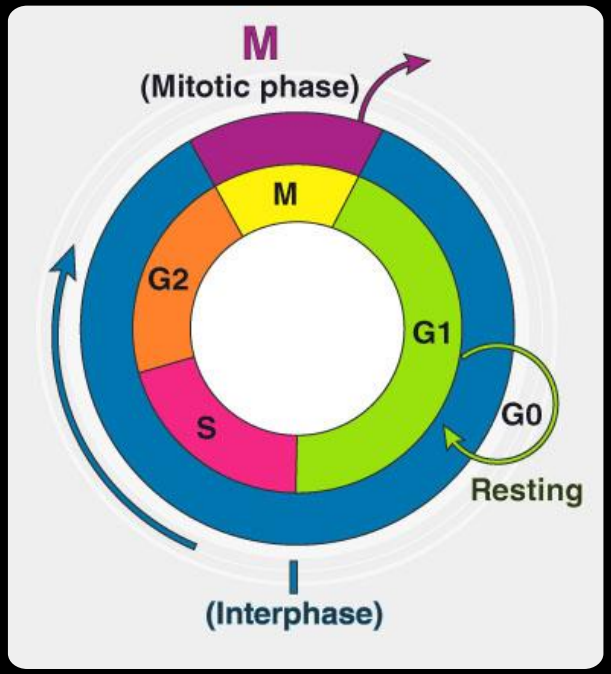
KEGG

Enrichment

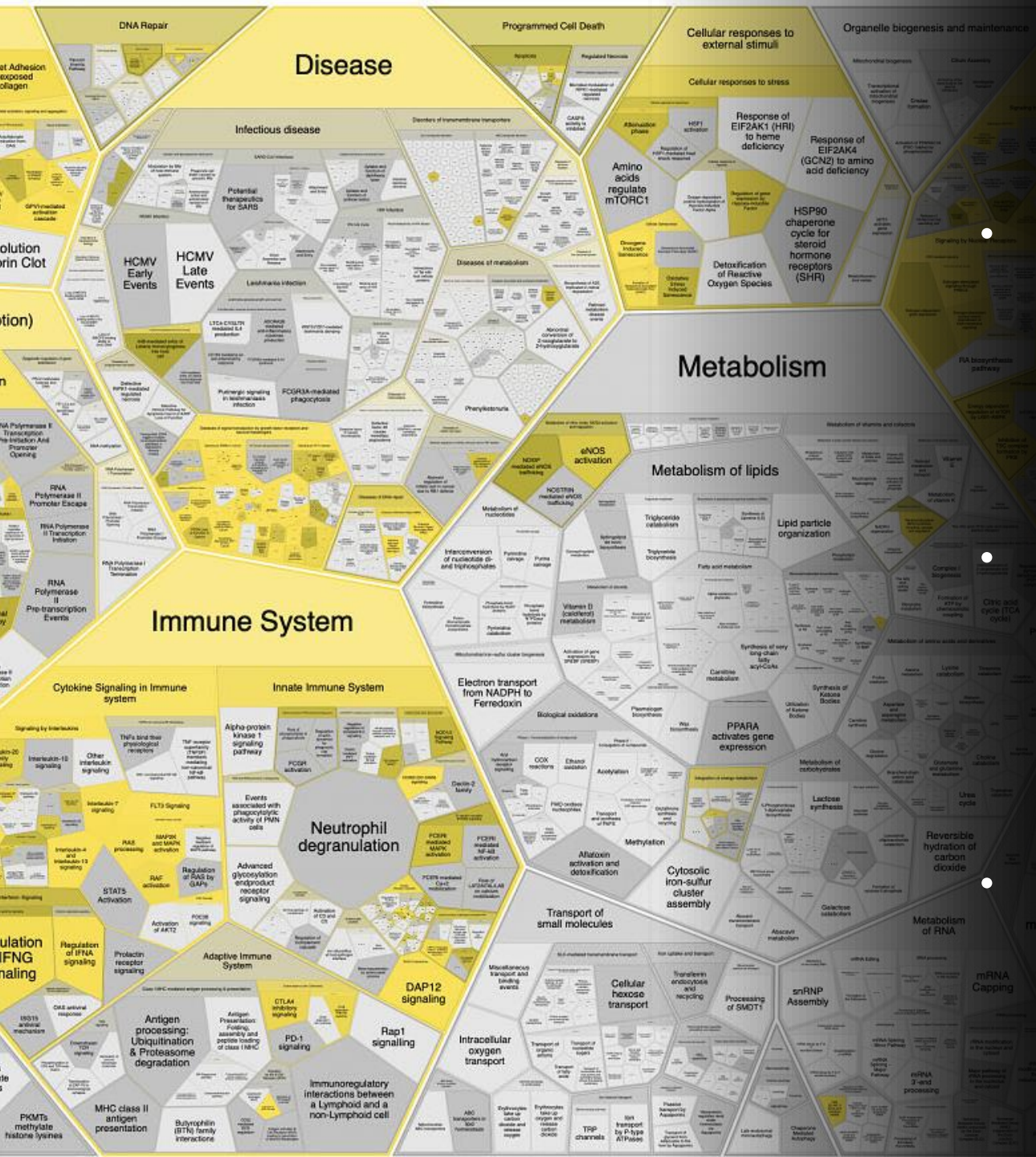
- Suppose a set of genes comes from a DE analysis between two experimental conditions C_1 and C_2.
  - Or from wherever, really.

- The set is "enriched" in a particular gene category if there are more genes from that category than we'd expect by chance.

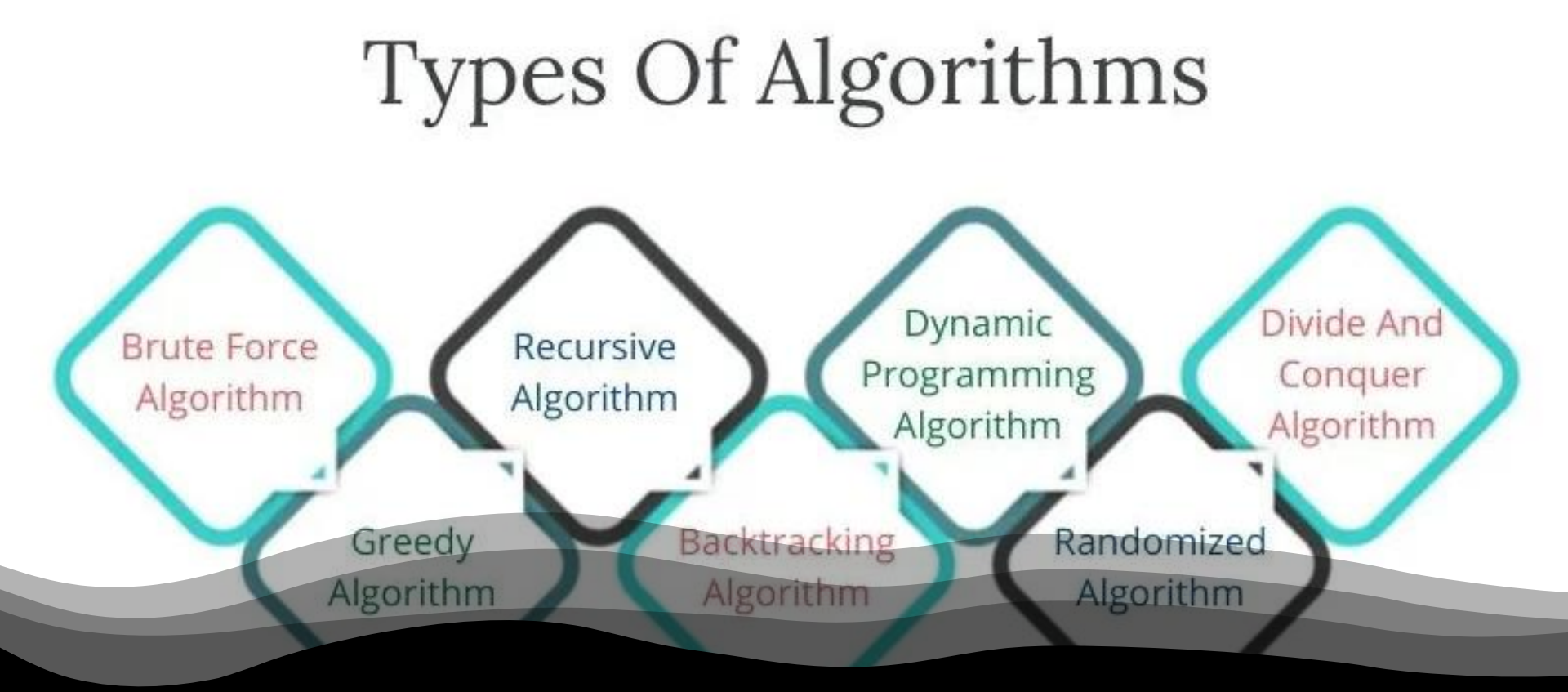
  - Suppose for example that only 1% of *all* genes are involved in the cell cycle.
    - But 80% of the DE list are cell cycle genes.
  - That (strongly) indicates the difference between C_1 and C_2 has something to do with the cell cycle.
    - Perhaps cells in C_1 are actively dividing and those in C_2 are quiescent.



Terminology



- People refer to this as "**Pathway Enrichment Analysis**"
  - Even if they are interested in all types of categories, not jut "pathways".
- It's also commonly called a "**GO Analysis**"
  - Even though "GO" is a particular collection of gene sets from 'The Gene Ontology Resource'.
- Also "**Gene Set Enrichment Analysis**"
  - Even though GSEA is a particular algorithm, it's also used generically.



Algorithms

- There are many algorithms that take a list of genes and determine which (if any) categories of genes are over-represented on the list.
- This is generally known as “Pathway Analysis”

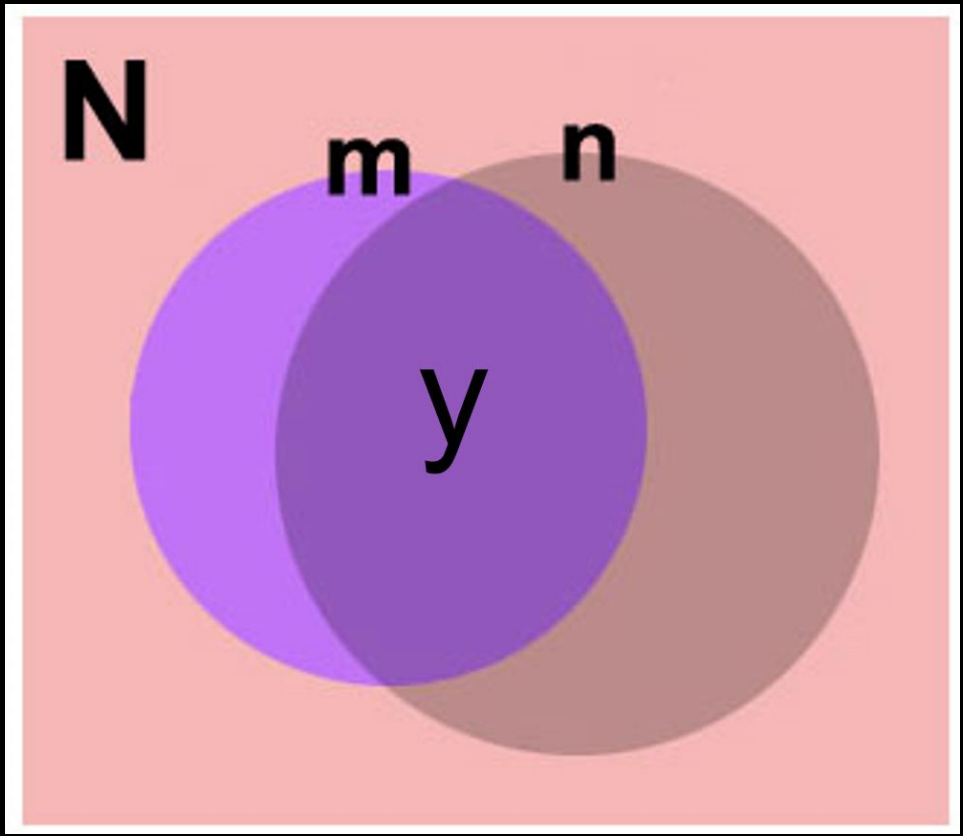
The Hypergeometric Test

- By far the most common and most straightforward approach is to model the problem by the hypergeometric distribution.
- As usual, this method makes a simplifying assumption.
  - For example, it assumes gene expression is independent between genes, which is definitely not true.

The Hypergeometric Random Variable

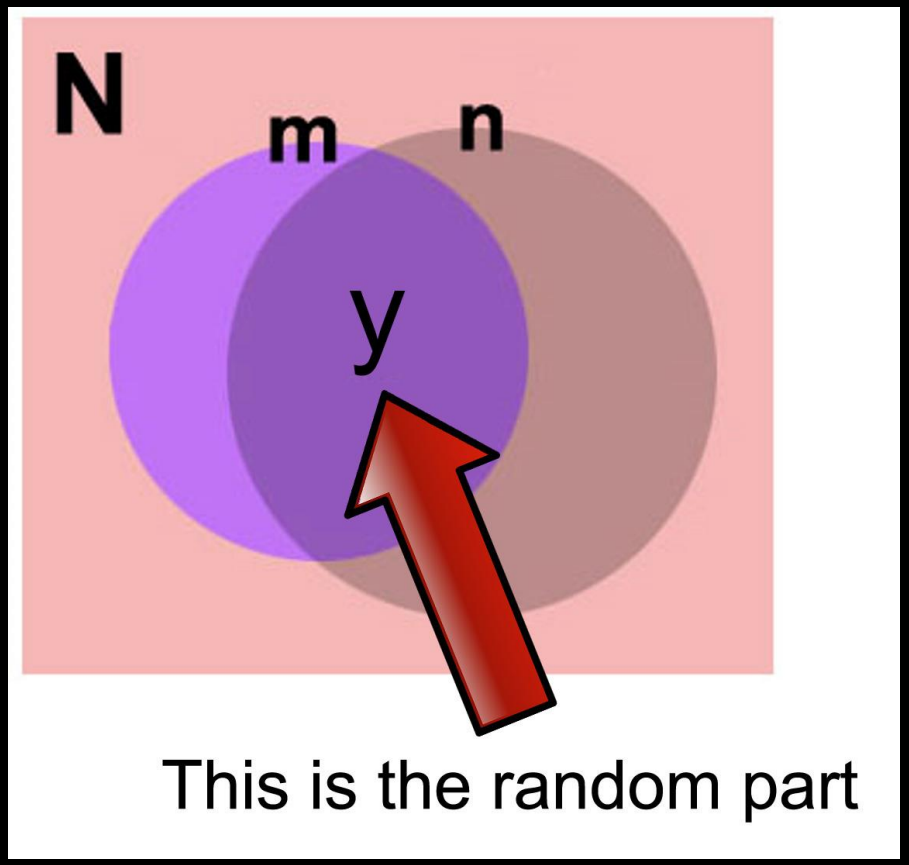
- Start with a set of *N* things.
- Randomly choose one subset of *m* things and another subset of *n* things.
- The observed value of the random variable is the size of the intersection, *y*.

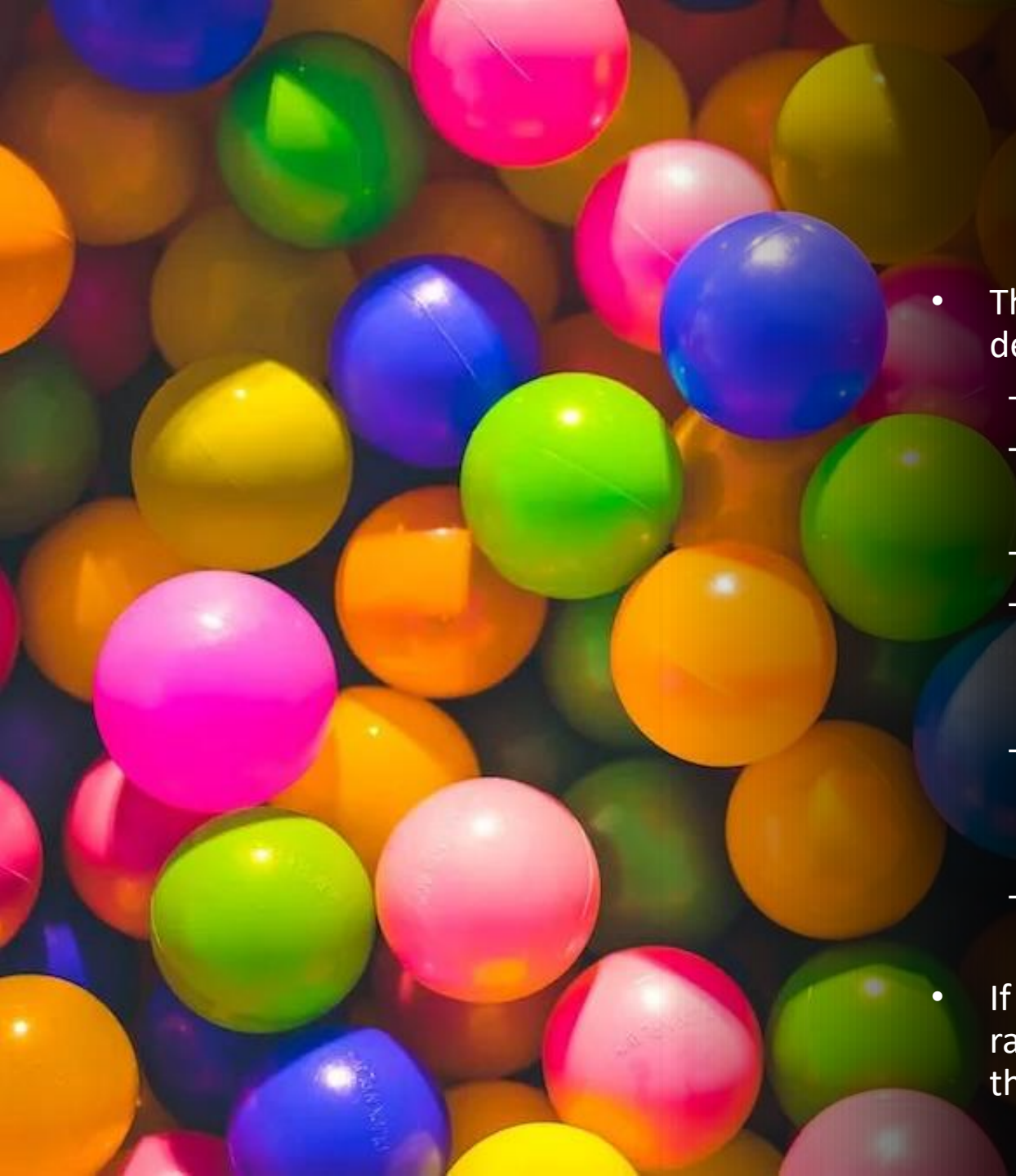
- *Note: If you look up (or google) the hypergeometric, you'll find it defined in a different (but equivalent) way.*
  - We are using this formulation because it's the best one for modeling pathway enrichment analysis.



y is random

- *N, m* and *n* are fixed.
- Each time we choose two subsets of sizes *m* and n, the size of their overlap *y* will vary.
- The probability distribution of y is called the "hypergeometric distribution".
  - We'll look at some graphs in a few slides.
  - It doesn't look much different from a normal, except with finite tails.





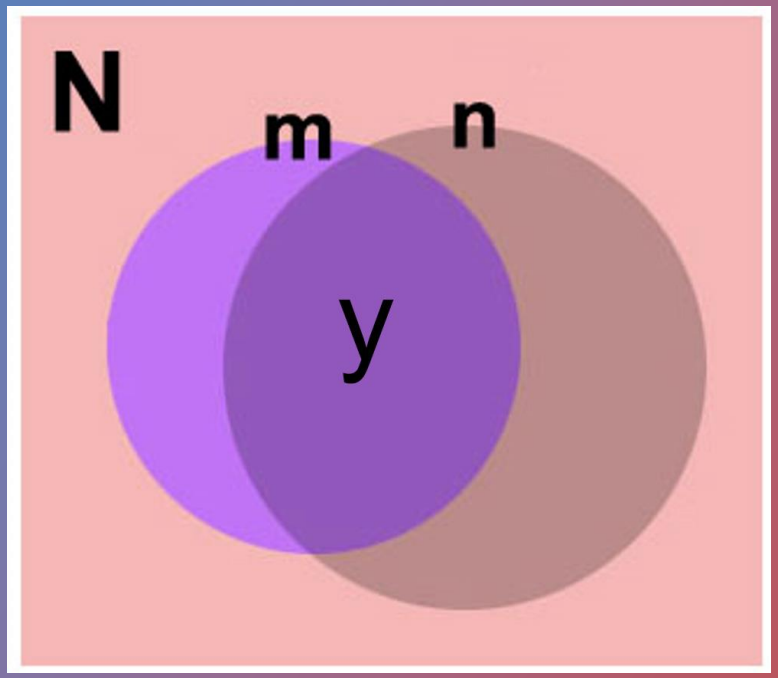
The Hypergeometric Distribution

- The hypergeometric distribution depends on three parameters.
  - Three positive integers.
  - *N*, *n* and *m*.

  - You have a set *S* of *N* things.
  - You select *n* of them at random.
    - Put them in a set S_1
  - You select *m* of them at random.
    - Put them in a set S_2
  - Let *Y* be the size of $S_1 \cap S_2$

- If we choose the sets S_1 and S_1 at random from the set of *N* things, then *Y* is a random variable.

p-values



- The distribution function of the hypergeometric random variable gives us $P(Y = y)$ where y is an observed value of the random variable.

$$P(Y = y) = \frac{\binom{N}{y}\binom{N-m}{n-y}}{\binom{N}{y}}$$

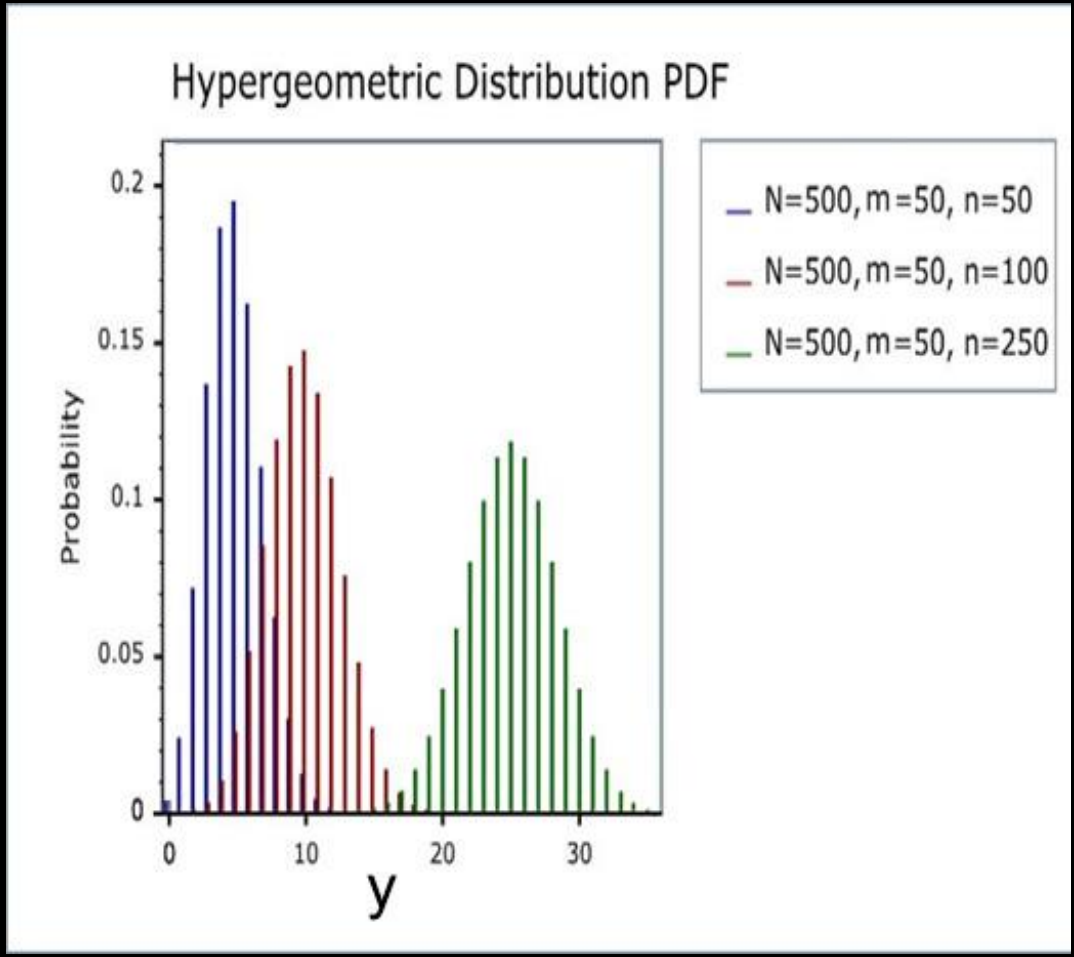
- The *p*-value is (as usual) the probability of an observed value *or more extreme*.

- For an observed intersection of size *k* the *p*-value is:

$$\sum_{i=k}^{\min\{m,n\}} P(Y = i)$$

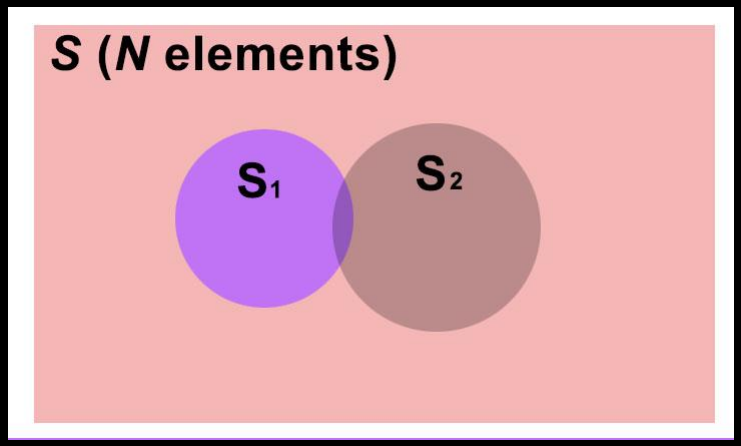
The Hypergeometric Distribution

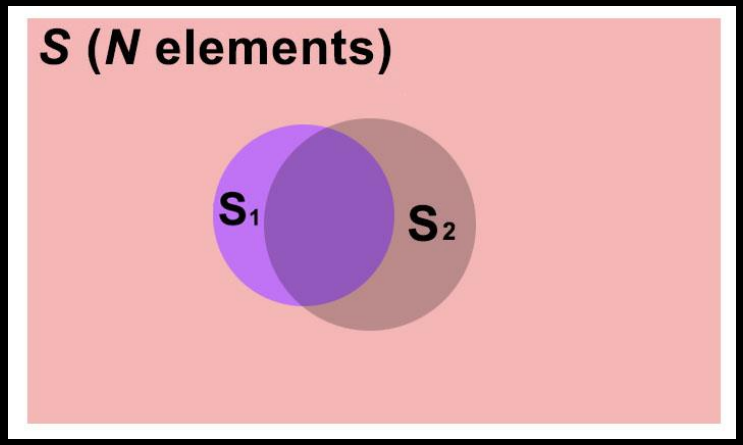
- Three Examples -



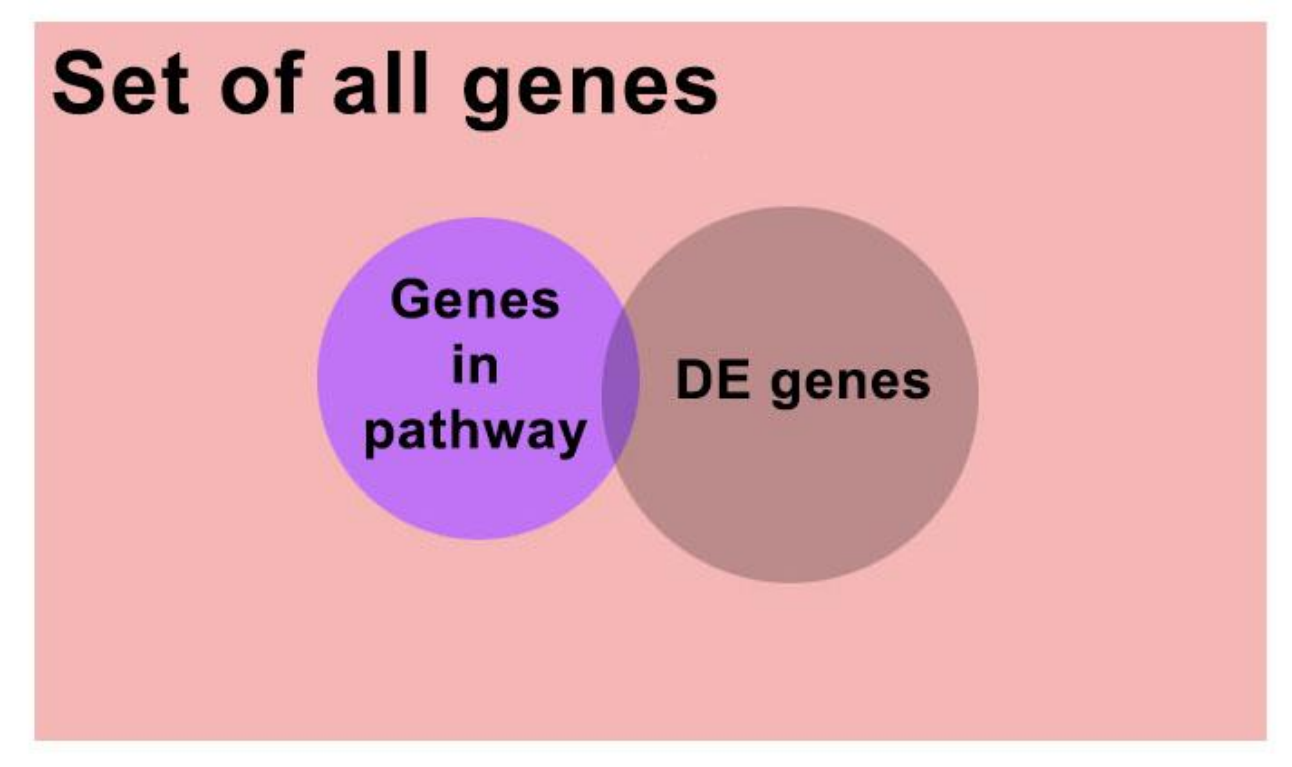
Intuition

- Here the intersection is about what you'd expect.
  - Overlap not significant.
  - Large hypergeometric *p*-value
- Here the intersection is higher than you'd expect.
  - Overlap significant
  - Small hypergeometric *p*-value.





Relation to Enrichment Analysis



- *A separate test is performed for each pathway.*
- The resultant *p*-values are then multiple-testing corrected to *q*-values, usually by Benjamini-Hochberg.

Example Results Table

- The input was a set of DE genes in an RNA-Seq experiment

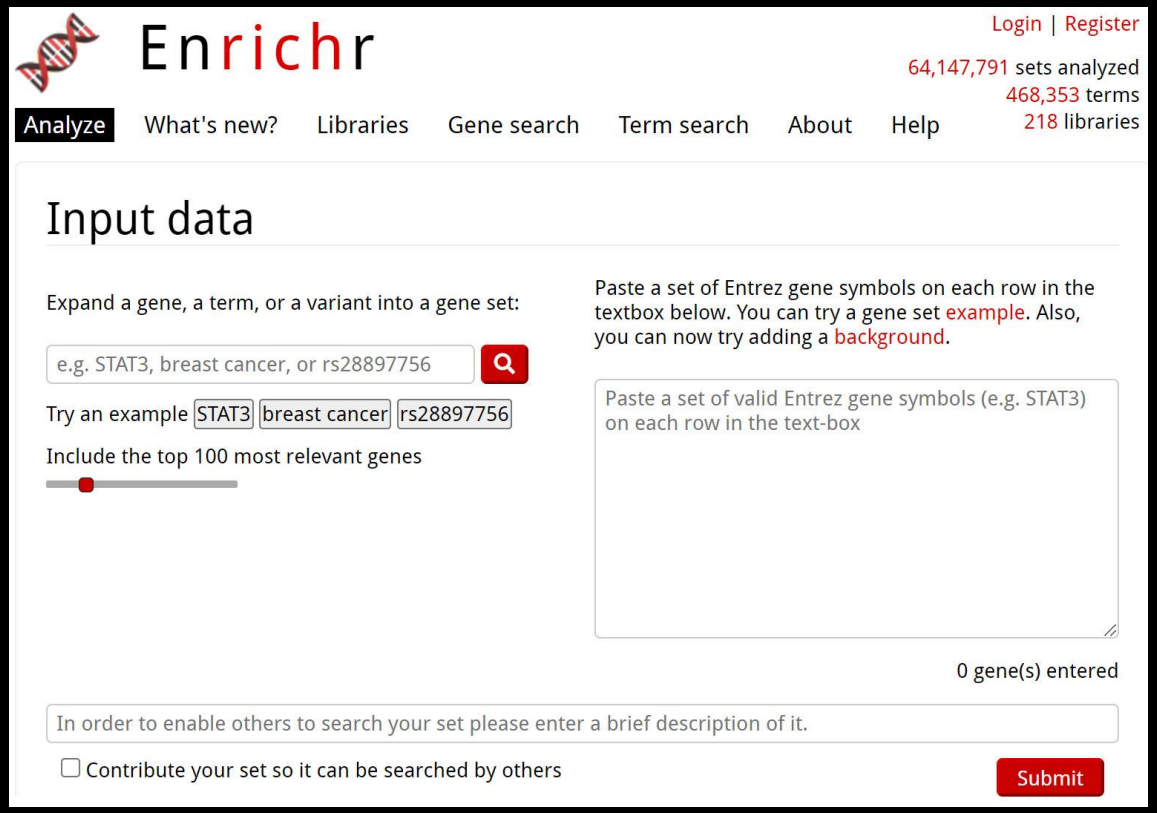
Category	Term	RT	Genes	Count	%	P-Value	Benjamini
GOTERM_CC_DIRECT	cytosol	RT		64	32.5	3.8E-6	1.1E-3
GOTERM_CC_DIRECT	nucleus	RT		82	41.6	7.3E-5	1.1E-2
GOTERM_MF_DIRECT	aminoacyl-tRNA ligase activity	RT		5	2.5	4.6E-4	8.7E-2
GOTERM_MF_DIRECT	protein kinase binding	RT		15	7.6	4.8E-4	8.7E-2
GOTERM_BP_DIRECT	translation	RT		11	5.6	6.2E-4	7.2E-1
GOTERM_CC_DIRECT	cytosolic ribosome	RT		6	3.0	7.1E-4	7.0E-2
GOTERM_CC_DIRECT	cytoskeleton	RT		25	12.7	9.6E-4	7.0E-2
GOTERM_MF_DIRECT	protein binding	RT		67	34.0	9.7E-4	1.2E-1
GOTERM_BP_DIRECT	tRNA aminoacylation for protein translation	RT		4	2.0	3.6E-3	1.0E0
GOTERM_MF_DIRECT	hydrolase activity, acting on glycosyl bonds	RT		5	2.5	5.0E-3	4.1E-1
GOTERM_CC_DIRECT	nucleoplasm	RT		45	22.8	5.4E-3	2.3E-1
GOTERM_MF_DIRECT	aminoacyl-tRNA editing activity	RT		3	1.5	5.6E-3	4.1E-1
GOTERM_BP_DIRECT	metabolic process	RT		7	3.6	6.3E-3	1.0E0
GOTERM_CC_DIRECT	Golgi apparatus	RT		23	11.7	6.6E-3	2.3E-1
GOTERM_CC_DIRECT	histone deacetylase complex	RT		4	2.0	6.6E-3	2.3E-1
GOTERM_CC_DIRECT	ribosome	RT		7	3.6	7.4E-3	2.3E-1
GOTERM_CC_DIRECT	membrane	RT		76	38.6	7.4E-3	2.3E-1
GOTERM_CC_DIRECT	polysome	RT		4	2.0	8.0E-3	2.3E-1
GOTERM_BP_DIRECT	cytoplasmic translation	RT		5	2.5	8.2E-3	1.0E0
GOTERM_CC_DIRECT	cytosolic small ribosomal subunit	RT		4	2.0	8.9E-3	2.4E-1
GOTERM_CC_DIRECT	microtubule organizing center	RT		6	3.0	9.7E-3	2.4E-1
GOTERM_BP_DIRECT	regulation of translation	RT		6	3.0	1.1E-2	1.0E0
GOTERM_BP_DIRECT	cellular response to epidermal growth factor stimulus	RT		4	2.0	1.2E-2	1.0E0
GOTERM_CC_DIRECT	trans-Golgi network	RT		7	3.6	1.3E-2	2.8E-1
GOTERM_BP_DIRECT	carbohydrate metabolic process	RT		7	3.6	1.3E-2	1.0E0
GOTERM_CC_DIRECT	cell projection	RT		19	9.6	1.4E-2	3.0E-1
GOTERM_CC_DIRECT	endoplasmic reticulum	RT		24	12.2	1.5E-2	3.0E-1
UP_KW_DOMAIN	Zinc-finger	RT		23	11.7	1.7E-2	3.0E-1
GOTERM_MF_DIRECT	RNA binding	RT		16	8.1	1.7E-2	8.3E-1
GOTERM_MF_DIRECT	valine-tRNA ligase activity	RT		2	1.0	1.8E-2	8.3E-1

Enrichment Analysis Free Servers

- There are several online resources for pathway enrichment analysis.
  - All of which seem to be based on the hypergeometric test.
- **DAVID**
  - U.S. Government
  - https://david.ncifcrf.gov/gene2gene.jsp
- **ENRICHR**
  - Mount Sinai
  - https://maayanlab.cloud/Enrichr/
- **STRING**
  - Has basically every species
  - EMBL, Europe (EMBL, Swiss, Denmark)
  - https://string-db.org/cgi/input?input_page_active_form=multiple_identifiers

ENRICHR

The interface is simple enough, just paste in your gene identifiers.



Speaking of Identifiers

- It specifically asks for Entrez gene symbols.
  - But the list might come to you as ENSEMBL ID's, RefSeq ID's, UCSC ID's, GENCODE ID's, etc.
  - So, you might have to do some ID conversion first.
- ENRICHR only works for mouse and human.
- DAVID is more flexible, but also pickier and buggier.

ENRICHR Example

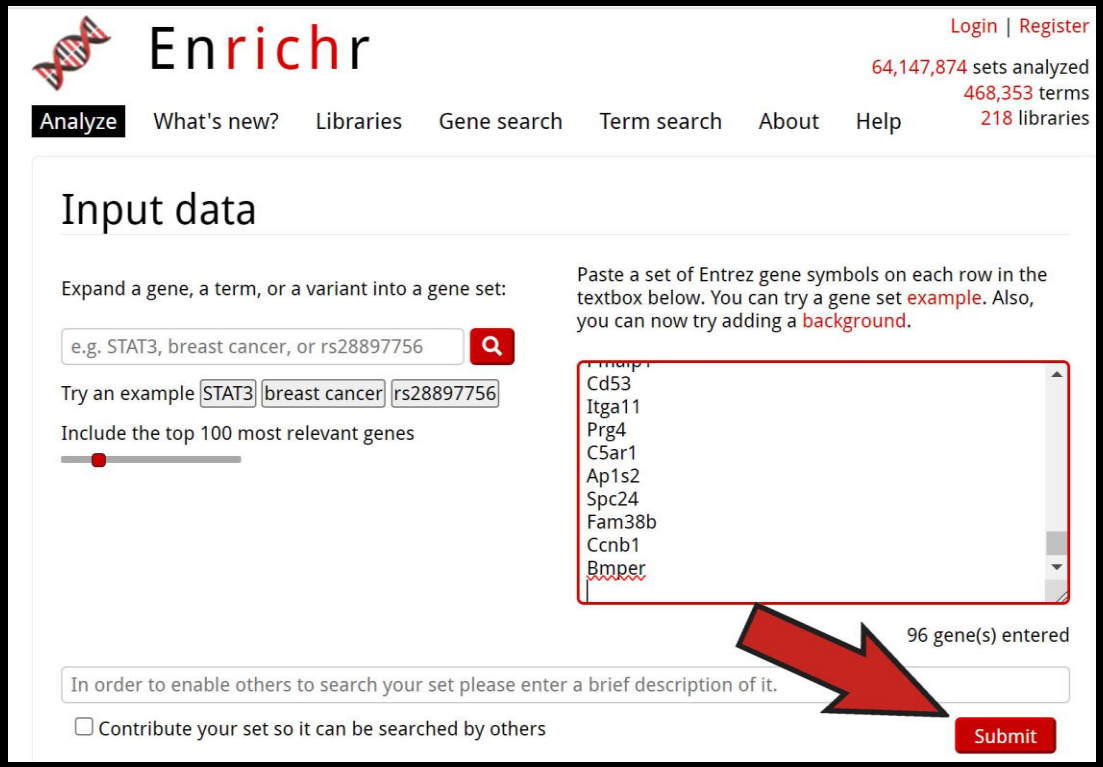
- This is the DE results from some old human fibroblast microarray data.

-bash

```
Upregulated Genes
-----------------
ID        Fold-Change     q-val
-----------------------------
Spp1      1.8729  0.015
Cxcr4     1.3826  0.015
Timp1     1.5253  0.015
Runx1     1.4677  0.015
Clec4n    1.762   0.02
Mmp3      1.3509  0.022
Dclk1     1.3832  0.025
Cthrc1    1.4832  0.027
Top2a     1.5814  0.027
2810417H13Rik     1.6237  0.028
Serpina3n         1.3107  0.028
Mmp12     1.7122  0.028
Casc5     1.616   0.028
Hells     1.5896  0.028
P4ha3     1.5362  0.028
Bcl2a1d   1.5912  0.028
Cenpe     1.5139  0.028
Tpx2      1.3353  0.028
Cks2      1.4676  0.028
Hmmr      1.5619  0.028
Cks2      1.5507  0.028
Kif11     1.5585  0.028
Ms4a7     1.5927  0.028
Cks2      1.5771  0.028
Ccna2     1.3341  0.028
Nuf2      1.5068  0.028
Bcl2a1b   1.6797  0.029
Ncapg2    1.2853  0.029
Mmp14     1.3202  0.029
Sh2d1b1   1.8439  0.029
Slc7a11   1.3913  0.029
Ccr2      1.4571  0.029
Pak1      1.3172  0.029
Msr1      1.4585  0.029
Mki67     1.464   0.029
:
```

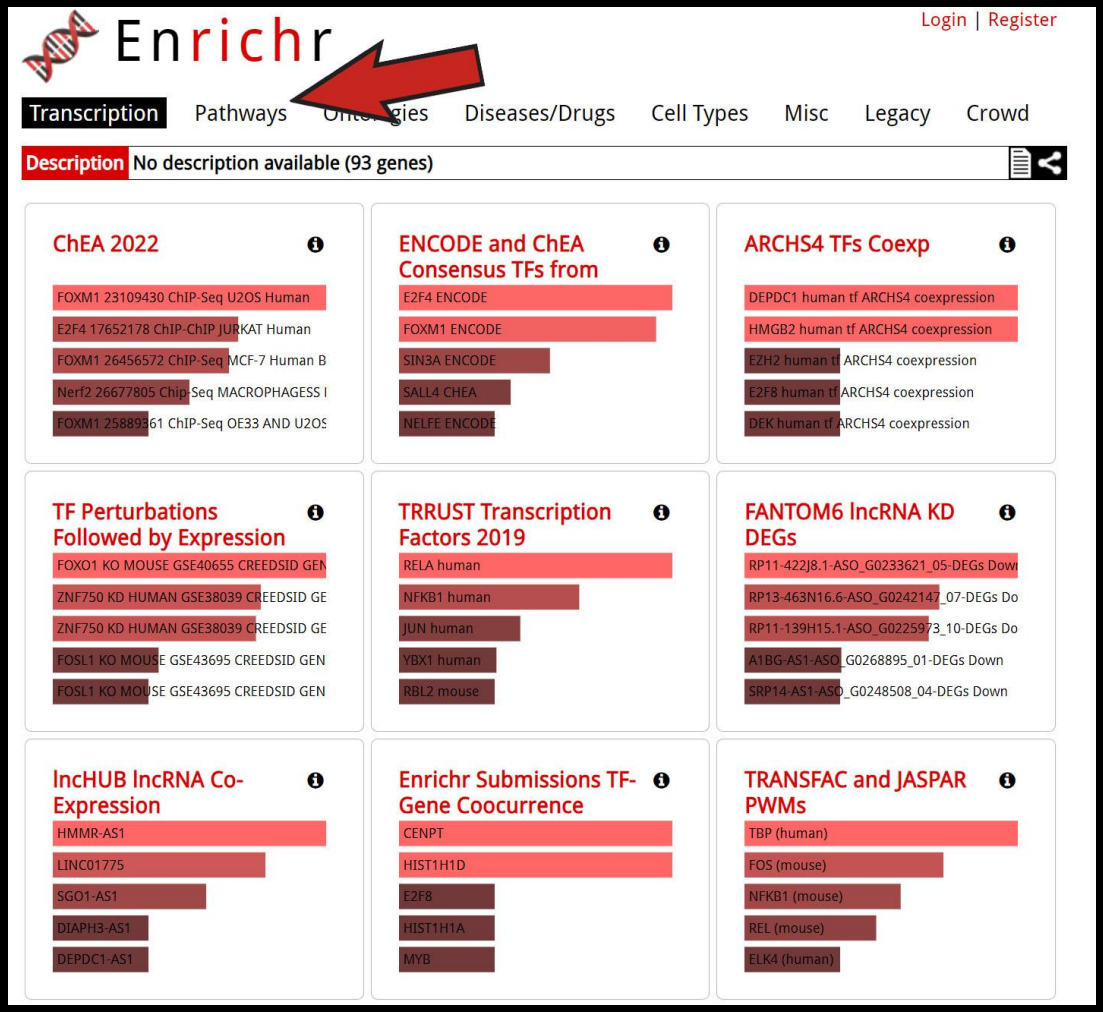
ENRICHR Example

- Paste in the top 100 genes and hit Submit



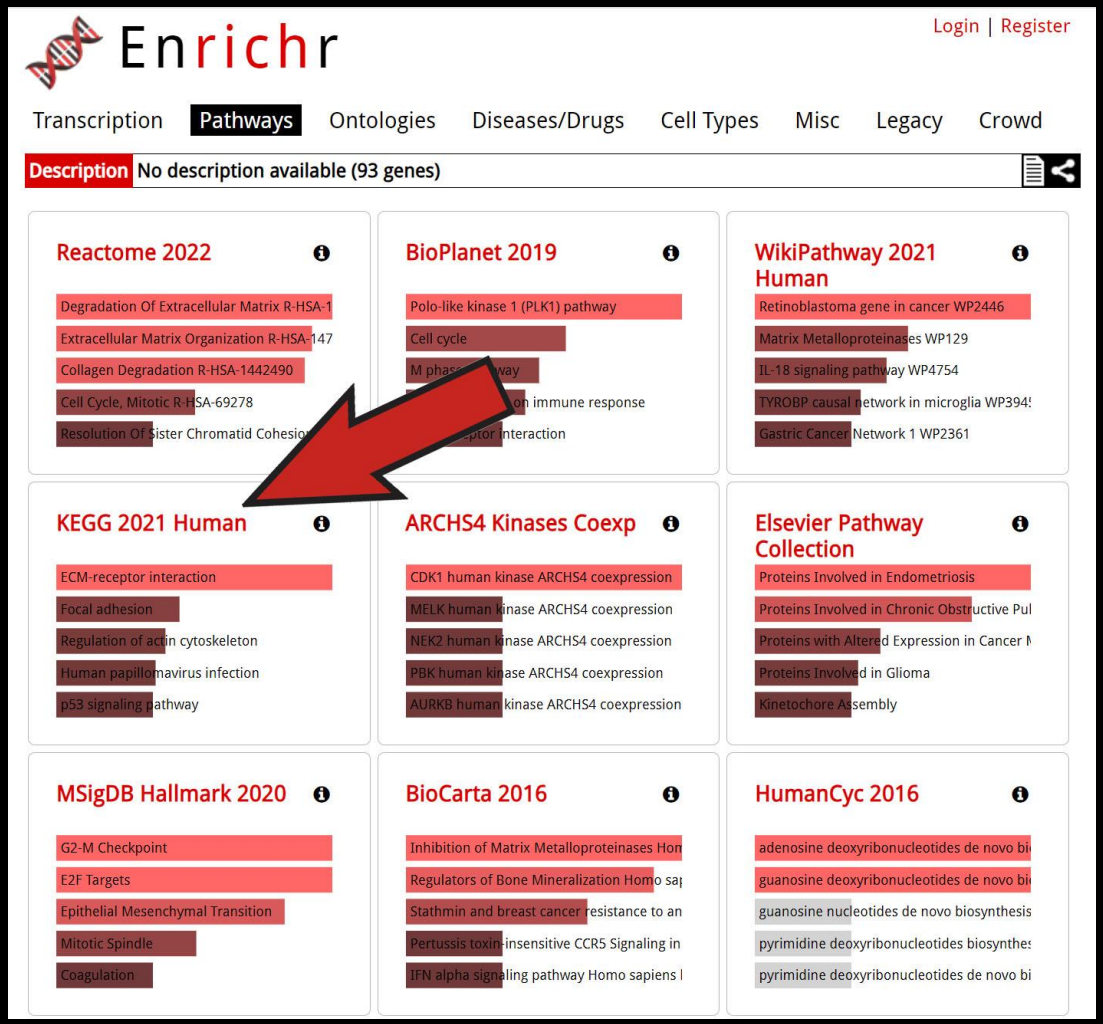
ENRICHR Example

- From this page there are several categories of gene sets available.
- Select Pathways.



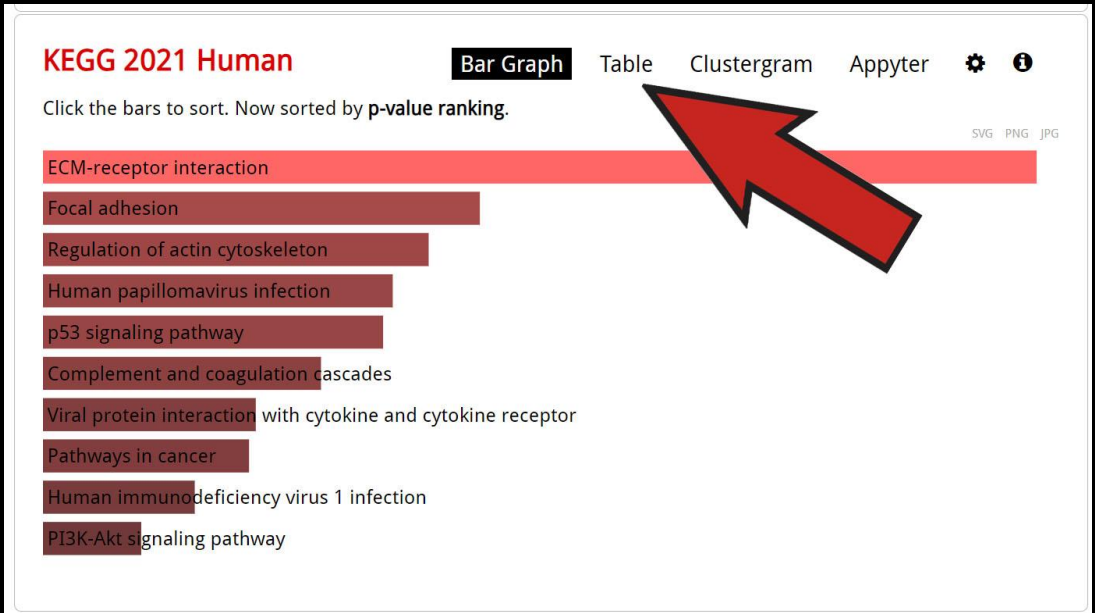
ENRICHR Example

- They have included a large number of pathway. KEGG is a fairly popular one.
  - Kyoto Encyclopedia of Genes and Genomes



ENRICHR Example

- This shows the basic KEGG summary.
- Click on Table.



ENRICHR Example

- "Adjusted *p*-value" here means "*q*-value".
- There's one highly significant and a bunch of others with *q*-value in the 0.1 range.
- There are also several pages of less significant pathways.

KEGG 2021 Human Bar Graph **Table** Clustergram Appyter

Hover each row to see the overlapping genes.

10 entries per page Search:

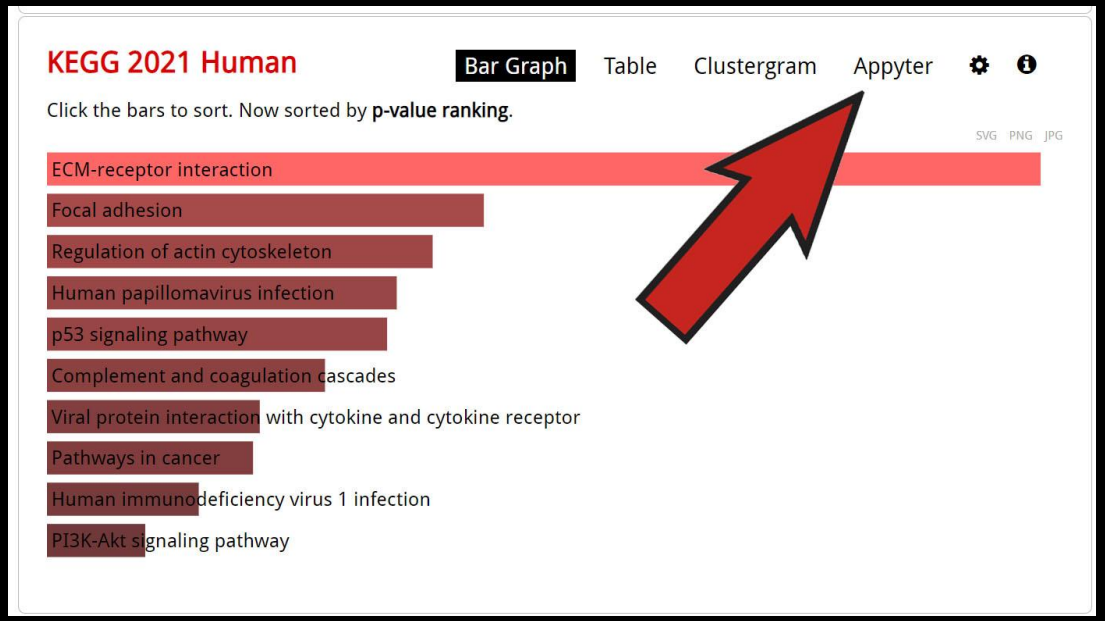
Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	ECM-receptor interaction	0.00005635	0.006142	13.57	132.77
2	p53 signaling pathway	0.004786	0.1043	9.45	50.46
3	Focal adhesion	0.002480	0.1043	5.71	34.28
4	Regulation of actin cytoskeleton	0.003513	0.1043	5.25	29.69
5	Human papillomavirus infection	0.004483	0.1043	4.16	22.47
6	Complement and coagulation cascades	0.007302	0.1326	8.06	39.65
7	Viral protein interaction with cytokine and cytokine receptor	0.01137	0.1622	6.81	30.47
8	Pathways in cancer	0.01190	0.1622	3.01	13.34
9	Human immunodeficiency virus 1 infection	0.01722	0.2086	4.26	17.29
10	Pyrimidine metabolism	0.02800	0.2628	8.08	28.89

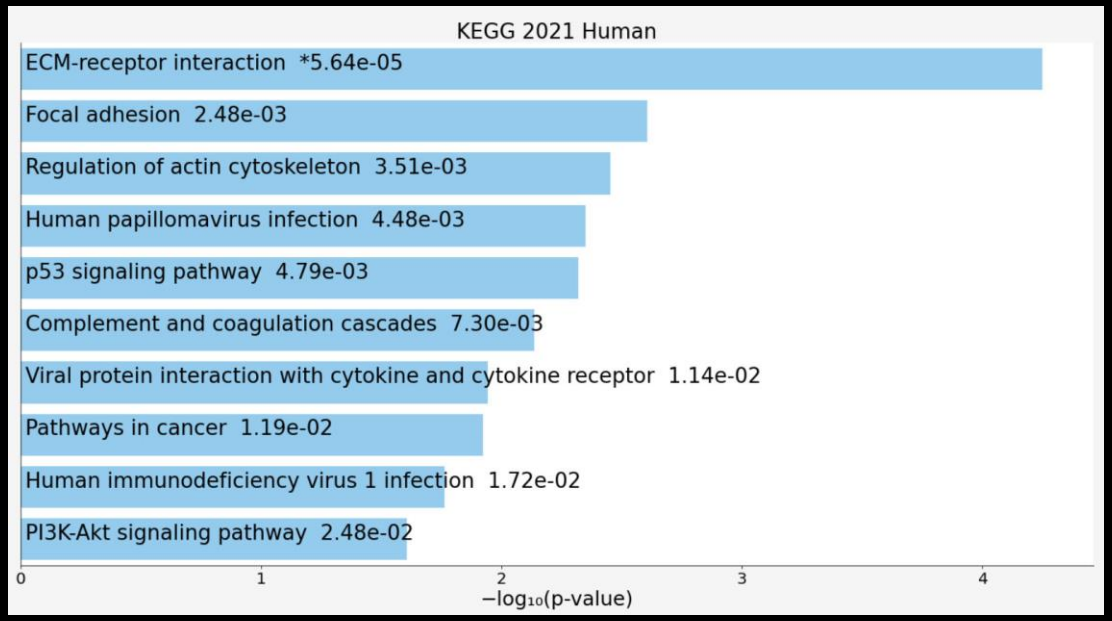
Showing 1 to 10 of 109 entries | Export entries to table Previous Next

Terms marked with an * have an overlap of less than 5

ENRICHR Example

- Click on Apptyer to get nicer visuals.





Background

- The bigger the background, the less likely subsets are to intersect.
- Therefore, if the background contains genes we didn't bother to measure, then the *p*-values will be artificially smaller than they should be.
- This is particularly a problem when working with microarray data.
  - In RNA-Seq we tend to measure all (or most) genes.
  - But there are reasons it can still be limited.
  - For example, we depend on one set of annotations for quantification that may not have all genes in the default background.
- ENRICHR does *not* let you set your own background.
  - DAVID does and so does Ingenuity.

Ingenuity Pathway Analysis (IPA)

Ingenuity is a private effort

- Spun off of Stanford actually.

They charge a lot of money for their product.

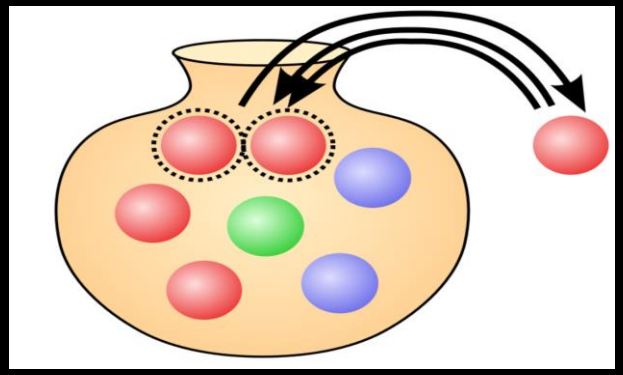
- Their justification is that they claim to have a large team of scientists updating their gene sets constantly.
- They claim to have the most up-to-date and curated gene sets.
- It might be true.

Arbitrary Choices

- We have to make an arbitrary choice of how many genes to include in the analysis.
  - We may only have exactly 200 significant genes with *q*-value < 1.
  - But typically, there are thousands of genes and many possible choices of *q*-value cutoff.
- When running Ingenuity, it can take an hour to get results back.
- So, people will typically try just one or a few cutoffs.

MPV

- Users have long desired a way to avoid this arbitrary choice.
  - It doesn't sound scientific to have to make such judgment calls.
- It would be nice to have a slider-bar you can move that updates the pathway analysis in real time as you move it.
  - But that's a huge amount of "on-the-fly" computation.
- But computation has caught up.
  - Here's a beta version we're working on in my lab.
  - DEMO MPV if there's time.



Hypergeometric Limitations

The hypergeometric test models the problem on independent draws of colored balls from an urn.

In this model, balls in the urn behave independently from each other.

But genes do not operate independently.

There can be two genes that are regulated together, so that they're both always at the same level.

In this case if one is DE then the other must also be DE.

Hypergeometric Assumptions

If we model two dependent genes as two independent balls in an urn, then we've acted like one thing is really two.

Whatever one does, the other must do.

But we're going to calculate *p*-values as if the same thing happened twice by chance when both either happen together or not, so it's really just one thing.

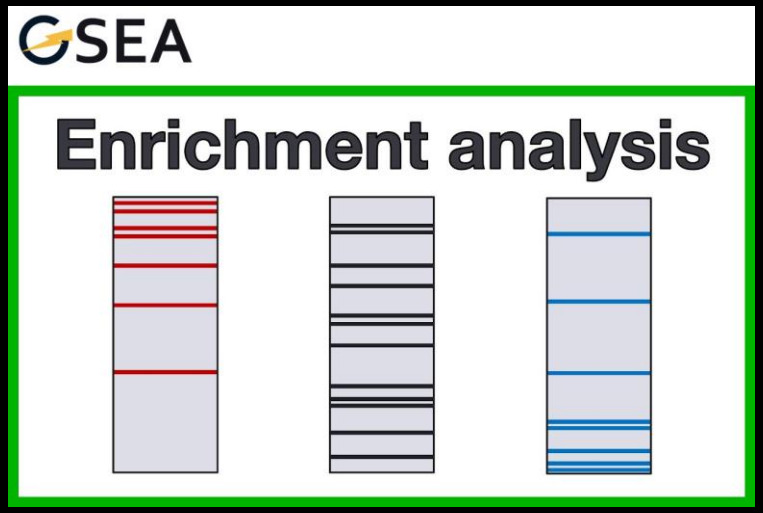
That will make things seem less likely than they really are, making *p*-values smaller than they should be, making us draw false conclusions.

Interpretation

- That's why (just like with BLAST) you should never get excited about a (hypergeometric) *p*-value or *q*-value on the order of 0.05 or even 0.01.
  - Even if there wasn't also a multiple testing issue.
- Therefore, we tend to disregard enrichment *q*-values unless they're particularly small.
  - Like 0.00001 or better yet something like 10E-127

GSEA

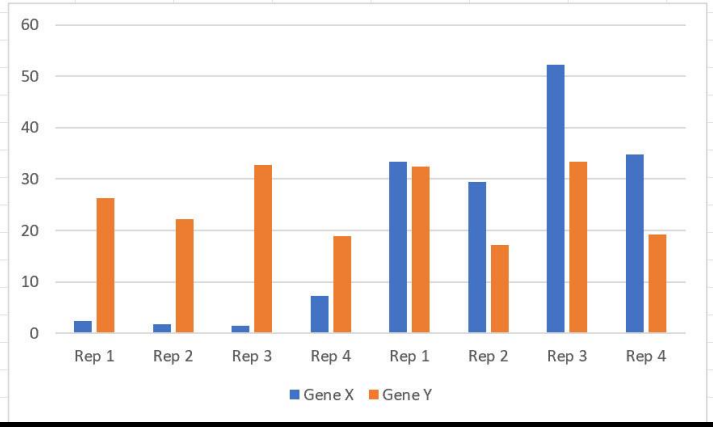
- GSEA is the name of an app that tries a different approach.
- Up to now we've been working with a subset of genes determined by a *q*-values cutoff.
- Instead GSEA works with all genes.
- First gene expression is correlated with the class label.
  - Class labels indicate experimental condition.
  - Example on next slide will help clarify what this means.



Expression/Class Correlation

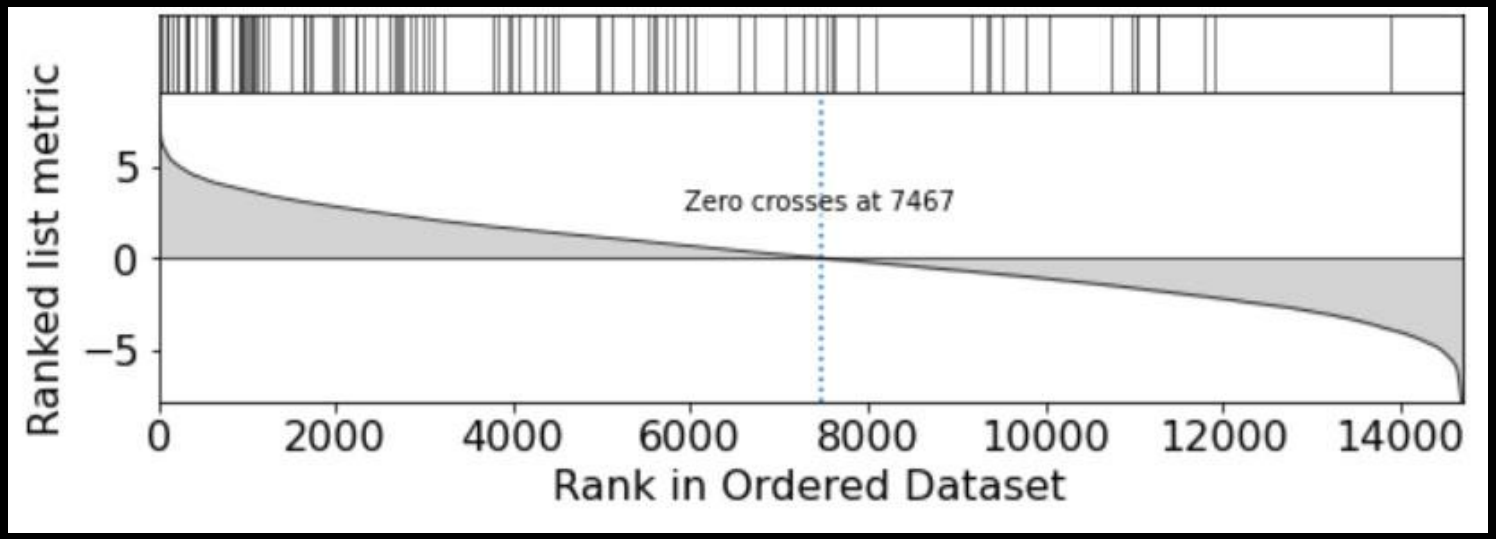
- This shows two genes and their correlation to the class labels, represented as a vector of 0's and 1's.
- There's no pathway in this story yet, that comes later.

	Condition 1				Condition 2				Correlation with Class Vector
	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2	Rep 3	Rep 4	
Gene X	2.4	1.8	1.4	7.2	33.4	29.42	52.17	34.83	**0.936986436**
Gene Y	26.3	22.24	32.75	18.91	32.4	17.2	33.4	19.23	**0.039800984**
Class	0	0	0	0	1	1	1	1	



GSEA Algorithm

- Next all genes are sorted by their correlations to the class labels.
- Then for each pathway, the genes in the pathway are indicated with dark lines.



- If the pathway has nothing to do with the differential expression between the conditions:
  - Then the dark lines should be randomly distributed uniformly across the x-axis.
- If the dark lines bunch up at either end (or even in the middle)
  - that indicates this gene set has something do with the difference between the conditions.

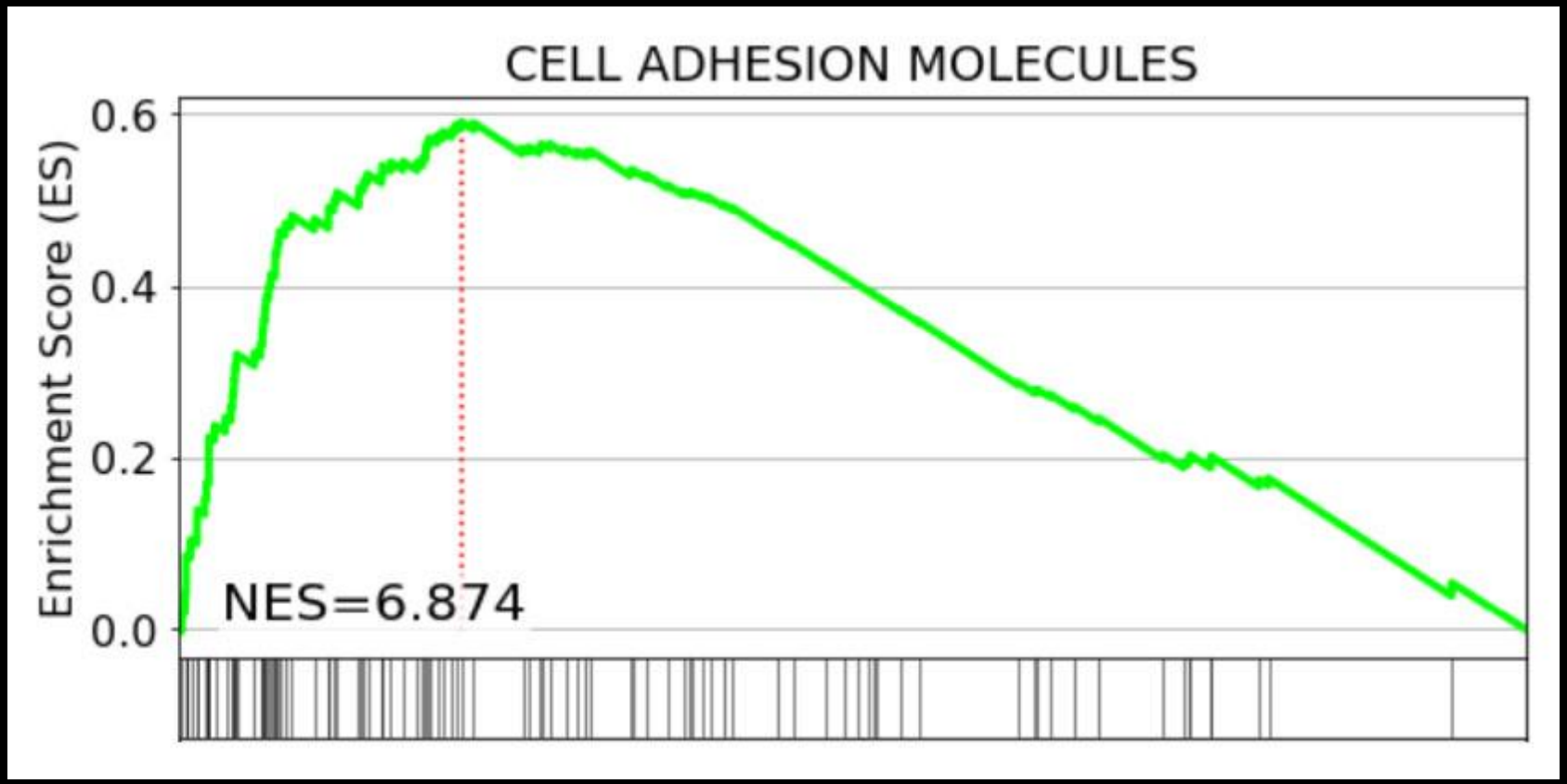
The GSEA Random Walk

- A random walk is then performed from left to right, going up at every gene in the pathway and down at every gene not in the pathway.
- Steps up are bigger than steps down.
  - This is from the GSEA paper

Calculate enrichment. We set the constant step size of the walk, so that it begins and ends with 0, and the area under the running sum is fixed to account for variations in gene set size. We walk down the list L, incrementing the running sum statistic by $\sqrt{(N - N_h)/N_h}$ when we encounter a gene in S and decrementing by $\sqrt{N_h/(N - N_h)}$ if the gene is not in S, where N is the number of genes in the list L, and N_h is the number of genes in the gene set S. The maximum deviation from zero is the ES for the gene set S, and corresponds to a standard Kolmogorov-Smirnov statistic.

The GSEA Enrichment Score

- The score for that gene set is then the maximum height achieved by the random walk.
  - Though a completely different problem, they were obviously inspired by BLAST
- Here the walk is represented by the green line.



GSEA Report

- A *p*-value is then calculated, which are then multiple-testing corrected for there being multiple gene sets.
- They use a non-parametric (permutation) approach to *p*-values, which is an upcoming topic.