Enrichment analysis with enrichR; what to be aware of

Bioinfo Seminar, DBM

Michal Kloc

Bentires-Alj Lab, University Hospital Basel, Switzerland, September 6, 2022

Outline

1. Overrepresentation analysis (ORA), mathematical formulation
2. enrichR: online tool for ORA
3. Critical evaluation of enrichR, pros and cons
4. Alternative: clusterProfiler

presentation inspired by
`https://www.pathwaycommons.org/guide/primers/statistics/fishers_exact_test/`

Overrepresentation analysis

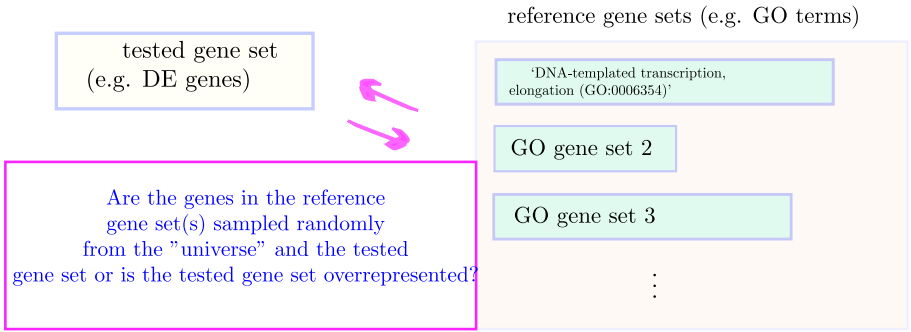
Task:

Check if a tested gene set shares an ‘unusual large number of genes’ with a reference gene set.

Overrepresentation analysis

Task:

Check if a tested gene set shares an 'unusual large number of genes' with a reference gene set.



Overrepresentation analysis

a simple case:

- 30 genes expressed in total (background)
- 15 genes are DE (tested gene set)
- 12 genes overlap with a reference gene set (transcription/elongation GO term)

Overrepresentation analysis

a simple case:

- 30 genes expressed in total (background)
- 15 genes are DE (tested gene set)
- 12 genes overlap with a reference gene set (transcription/elongation GO term)

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	12	3	15
NOT IN Transcription Elongation	3	12	15
Total	15	15	30

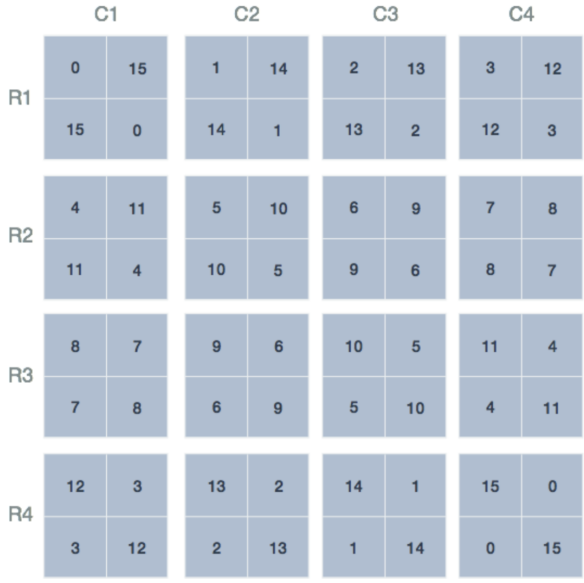
Fisher's exact test

null hypothesis H_0: The two gene sets are independent

Establish the p value for our observation.

Fisher's exact test

The number of possible configurations (contingency tables) is finite but not all are equally probable



Fisher's exact test

$$\mathcal{P} = \frac{\binom{15}{1}\binom{15}{14}}{\binom{30}{15}} = \frac{\binom{15}{1}^2}{\binom{30}{15}} = 1.45 \times 10^{-6}$$

Hypergeometric distribution

$$f(x; N, n, r) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \text{ for } x = 0, 1 \ldots r$$

x: random variable (overlap), N: background size, r: size of the tested gene set, n: size of the reference gene set

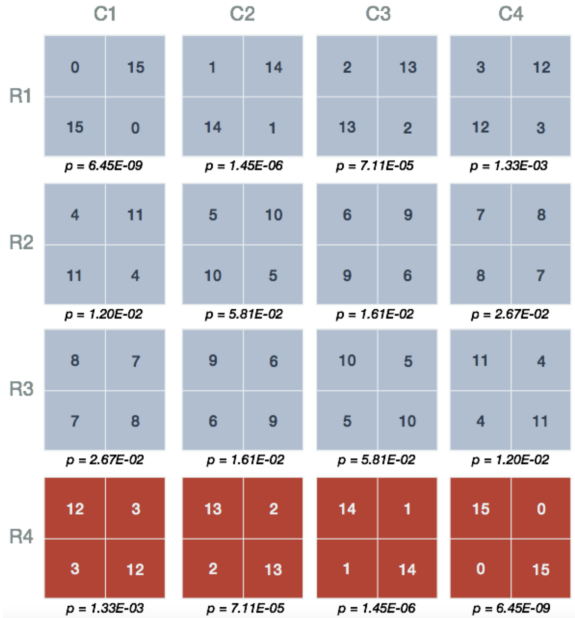
	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	1	14	15
NOT IN Transcription Elongation	14	1	15
Total	15	15	30

Fisher's exact test

One-sided (directional) test:

$$
\begin{aligned}
pval &= 1.33 \times 10^{-2} + 7.11 \times 10^{-5} \\
&\quad + 1.45 \times 10^{-6} + 6.45 \times 10^{-9} \\
&= 0.001407159
\end{aligned}
$$

probabilities:



Enrichr

Input data

Expand a gene, a term, or a variant into a gene set:

e.g. STAT3, breast cancer, or rs28897756

Try an example STAT3 breast cancer rs28897756

Include the top 100 most relevant genes

Paste a set of valid Entrez gene symbols on each row in the text-box below. Try a gene set example.

Paste a set of valid Entrez gene symbols (e.g. STAT3) on each row in the text-box

0 gene(s) entered

In order to enable others to search your set please enter a brief description of it.

Contribute your set so it can be searched by others

Submit

Enrichr

Login | Register

Transcription Pathways Ontologies Diseases/Drugs Cell Types Misc Legacy Crowd

Description No description available (234 genes)

BioPlanet 2019

- Beta-1 integrin cell surface interactions
- ECM-receptor interaction
- Collagen biosynthesis and modifying enzym
- Extracellular matrix organization
- Focal adhesion

WikiPathway 2021 Human

- miRNA targets in ECM and membrane recep
- Focal Adhesion WP306
- VEGFA-VEGFR2 Signaling Pathway WP3888
- Focal Adhesion-PI3K-Akt-mTOR-signaling pa
- Type I collagen synthesis in the context of O

KEGG 2021 Human

- Focal adhesion
- ECM-receptor interaction
- Protein digestion and absorption
- PI3K-Akt signaling pathway
- Proteoglycans in cancer

ARCHS4 Kinases Coexp

- DDR2 human kinase ARCHS4 coexpression
- PDGFRB human kinase ARCHS4 coexpressio
- PDGFRA human kinase ARCHS4 coexpressio
- RPS6KA2 human kinase ARCHS4 coexpressic
- MYLK human kinase ARCHS4 coexpression

Elsevier Pathway Collection

- Proteins with Altered Expression in Cancer M
- Proteins Involved in Glioblastoma
- Invadopodia Formation in Cancer Cells
- Glioma Invasion Signaling
- Synovial Fibroblast Proliferation in Rheumat

MSigDB Hallmark 2020

- Epithelial Mesenchymal Transition
- Apical Junction
- Coagulation
- Complement
- Angiogenesis

MSigDB Hallmark 2020

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	Epithelial Mesenchymal Transition	3.880e-51	1.707e-49	34.41	3993.73
2	Angiogenesis	1.715e-7	0.000001510	20.99	326.95
3	Apical Junction	2.478e-11	5.453e-10	8.97	218.98
4	Coagulation	8.946e-9	1.312e-7	9.24	171.29
5	Complement	1.052e-7	0.000001158	6.70	107.63
6	IL-2/STAT5 Signaling	6.680e-7	0.000004899	6.19	88.05
7	UV Response Dn	0.00005096	0.0003203	5.82	57.49
8	Myogenesis	0.0001284	0.0007060	4.60	41.22
9	Protein Secretion	0.0009280	0.004083	5.75	40.17
10	Apoptosis	0.0006305	0.003082	4.54	33.44

Showing 1 to 10 of 44 entries | Export entries to table

Previous Next

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

enrichR: not adjustable

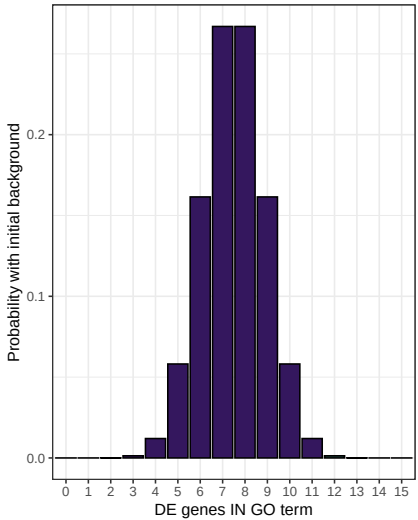
genes, which has advantages and disadvantages. Enrichr does not have an ID conversion tool, which is highly desired by many users. Enrichr also does not have the ability to upload a background list, and it does not have implementation of parametric tests such as Gene Set Enrichment Analysis (GSEA) (40), Parametric Analysis of Gene set Enrichment (PAGE) (9), and our own Principal Angle Enrichment Analysis (PAEA) (41). These features are planned.

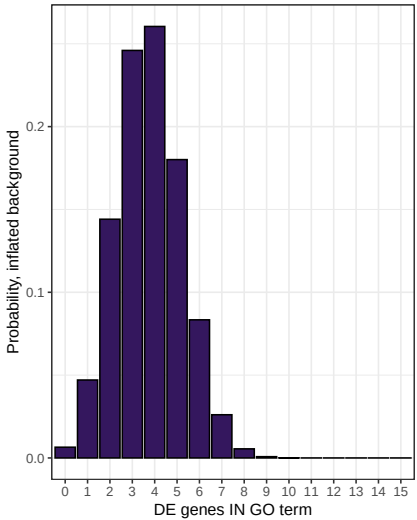
Effect of the background

```r
x <- c(0:15)
N <-  30 #universe
r <- 15 #DE genes (success)
n <- 15 #reference gene set

probabilities <- dhyper(x, r, N - r, n, log = FALSE)
pvalue <- sum(probabilities[13:16])
pvalue
#[1] 0.001407165

probabilities.BG <- dhyper(x, r, 2*N - r, n, log = FALSE)
pvalue.BG <- sum(probabilities.BG[13:16])
pvalue.BG
#[1] 1.233422e-07
```

dhyper(x, m, n, k, log = FALSE)

Effect of the background





Alternative: clusterProfiler

calling the function in R

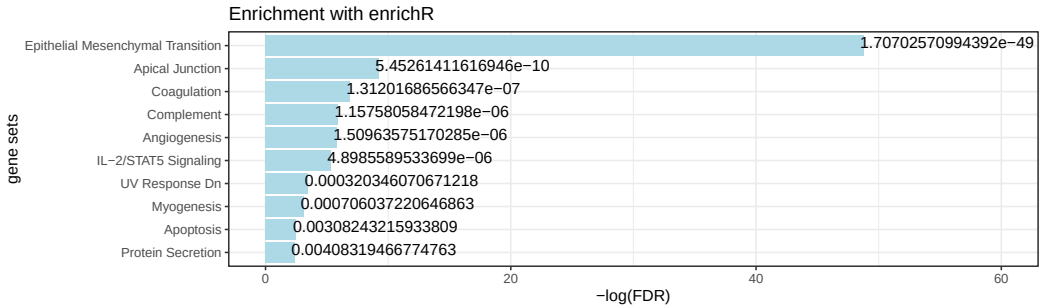
```
xx.2 <- compareCluster(genes.to.test.EntrezID, enricher, TERM2GENE=m_t2g.2,
                       pvalueCutoff=0.05, pAdjustMethod="BH")
```

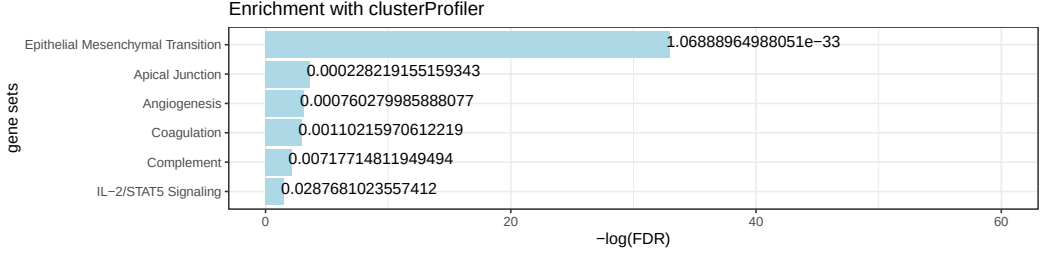
```
> head(m_t2g.2)
# A tibble: 6 × 2
  gs_name                           entrez_gene
  <chr>                             <chr>
1 TNF-alpha Signaling via NF-kB 17118
2 TNF-alpha Signaling via NF-kB 83430
3 TNF-alpha Signaling via NF-kB 18081
4 TNF-alpha Signaling via NF-kB 21950
5 TNF-alpha Signaling via NF-kB 17691
6 TNF-alpha Signaling via NF-kB 11910
```

By defauls, the background is formed by
the union of genes from the tested data sets

```
xx.3 <- compareCluster(genes.to.test.EntrezID, enricher, TERM2GENE=m_t2g.2,
                       pvalueCutoff=0.05, pAdjustMethod="BH",
                       universe = keys(mm))
```

Comparison


Enrichment with enrichR
Epithelial Mesenchymal Transition
Apical Junction
Coagulation
Complement
Angiogenesis
IL−2/STAT5 Signaling
UV Response Dn
Myogenesis
Apoptosis
Protein Secretion
1.70702570994392e−49
5.45261411616946e−10
1.31201686566347e−07
1.15758058472198e−06
1.50963575170285e−06
4.8985589533699e−06
0.000320346070671218
0.000706037220646863
0.00308243215933809
0.00408319466774763
gene sets
0
20
40
60
−log(FDR)


Enrichment with clusterProfiler
Epithelial Mesenchymal Transition
Apical Junction
Angiogenesis
Coagulation
Complement
IL−2/STAT5 Signaling
1.06888964988051e−33
0.000228219155159343
0.000760279985888077
0.00110215970612219
0.00717714811949494
0.0287681023557412
gene sets
0
20
40
60
−log(FDR)

Summary

- enrichR is quite widely used mainly for its userfriendliness
- Impressive collection of databases, updated
- Biggest problem: not adjustable (background!!!), apparently very broad
  $\rightsquigarrow$ false positives
- There are alternatives but not so easy to use. However, more contralable (clusterProfiler)
- Despite the test being exact, different implementations (packages) give somewhat different results

Summary