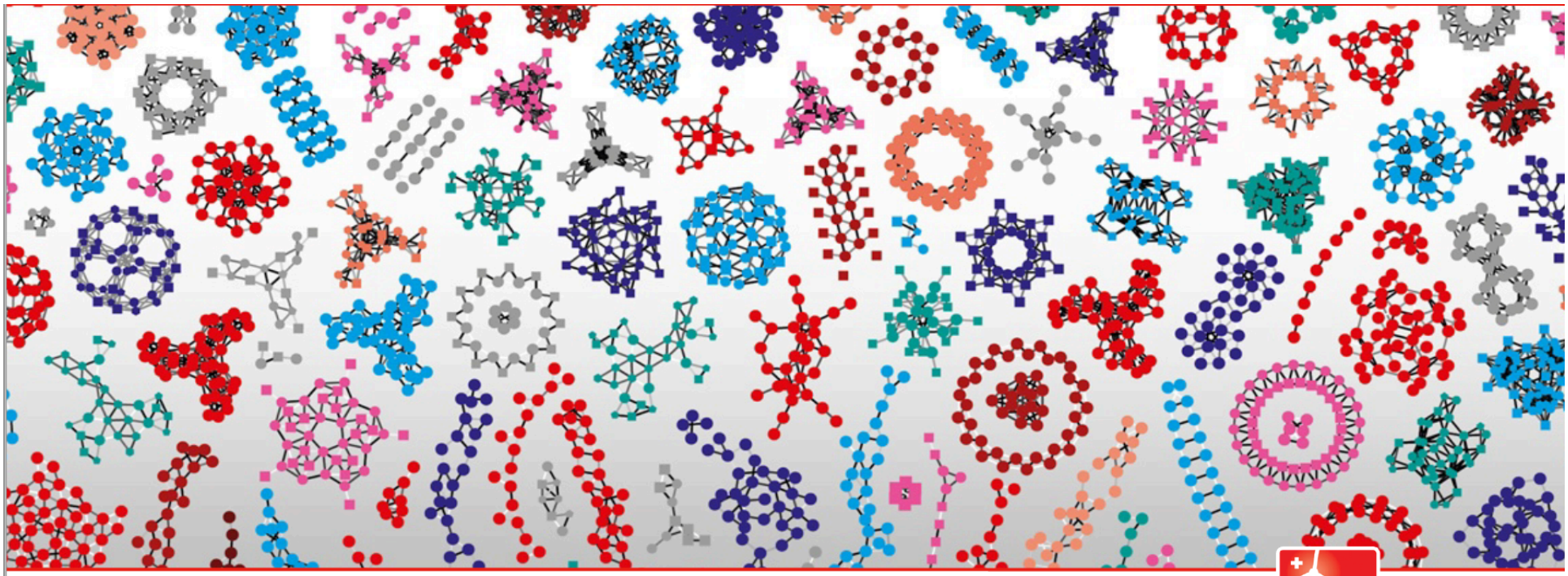


Swiss Institute of Bioinformatics

Single cell transcriptomics data analysis

Day 3

Differential gene expression and enrichment analysis

Differential gene expression (DGE) analysis

- FindAllMarkers(): implemented in Seurat, defaults to Wilcoxon test, to detect genes that are “markers” for cell clusters. Finds genes that are DE between 1 cluster and all other cells.

- FindMarkers(): to perform pairwise DGE analysis, eg between cluster 1 and cluster 2, defaults to Wilcoxon test.

What is the ideal DGE analysis method?

Bias, robustness and scalability in single-cell differential expression analysis

Charlotte Soneson[1,2] & Mark D Robinson[1,2]

Many methods have been used to determine differential gene expression from single-cell RNA (scRNA)-seq data. We evaluated 36 approaches using experimental and synthetic data and found considerable differences in the number and characteristics of the genes that are called differentially expressed. Prefiltering of lowly expressed genes has important

recent studies suggest that the optimal metl
the number of cells and strength of the signa
not initially developed for scRNA-seq analysi
In this study, we used processed data sets,
other sources, to evaluate DE methods in sc
study expands the number of methods and ra

https://www.nature.com/articles/nmeth.4612

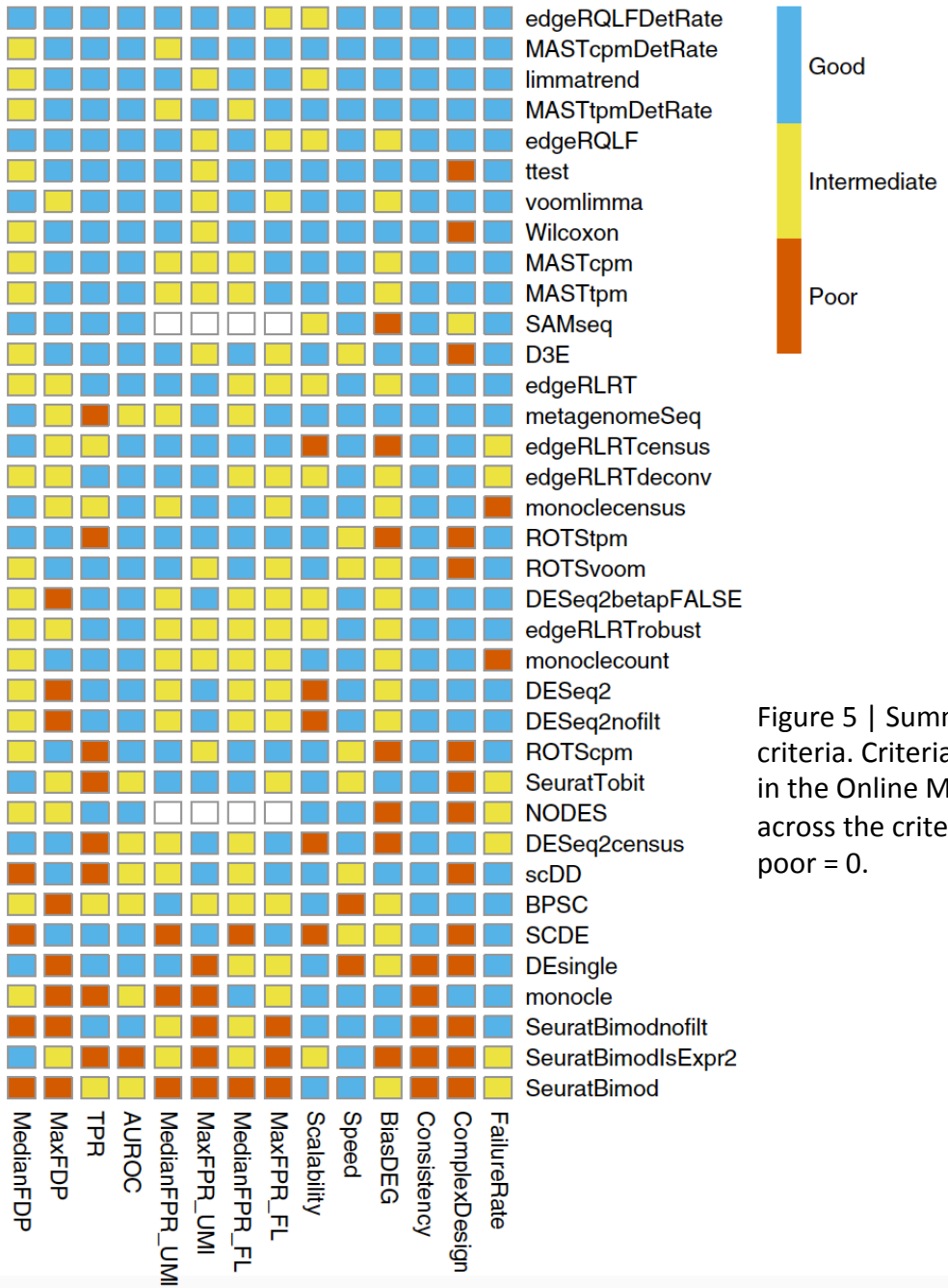


Figure 5 | Summary of DE method performance across all major evaluation criteria. Criteria and cutoff values for performance categories are available in the Online Methods. Methods are ranked by their average performance across the criteria, with the numerical encoding good = 2, intermediate = 1, poor = 0.

https://github.com/csoneson/conquer_comparison/blob/master/scripts/apply_edgeRQLFDetRate.R

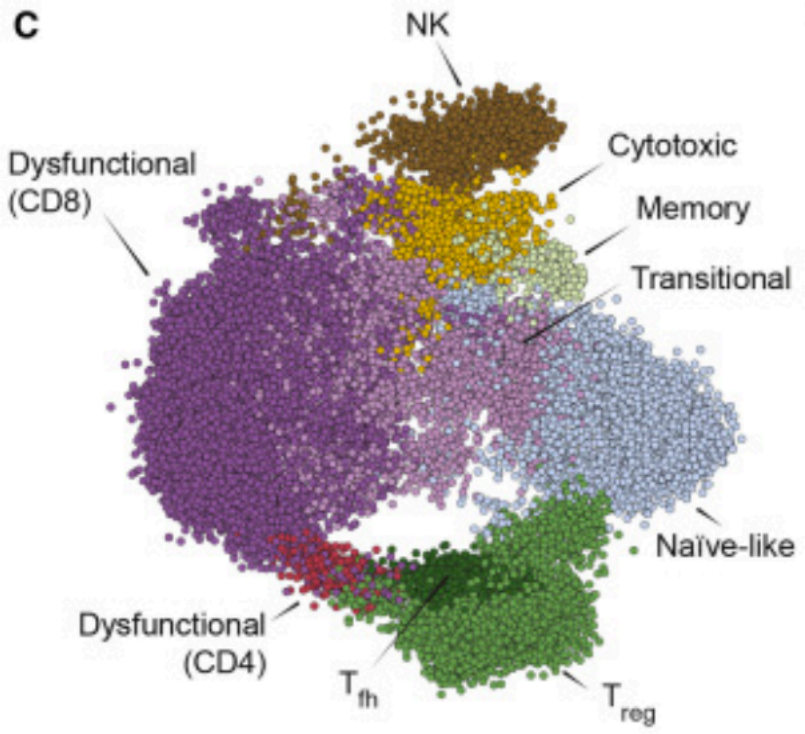
limma or edgeR

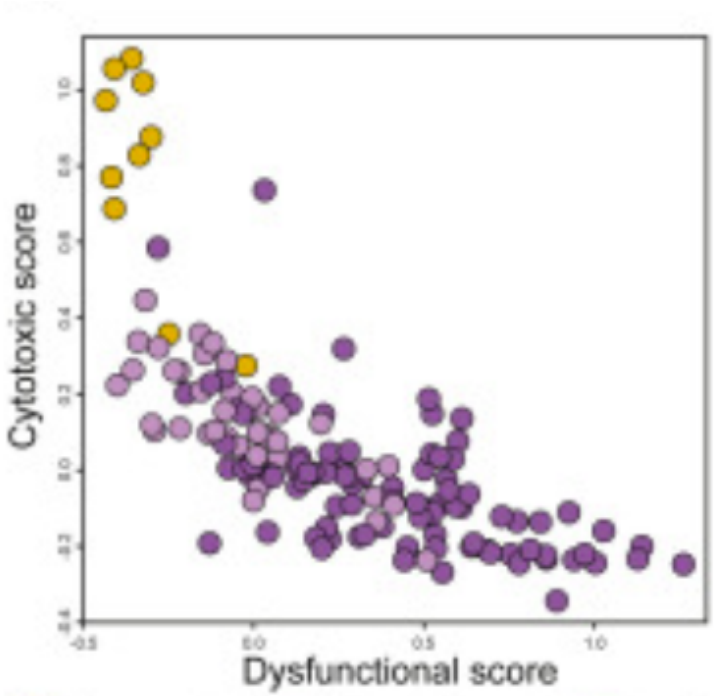
- Methods designed for bulk RNA seq analysis
- Can be used to include batch effects in model as covariates
- Can be used to analyze factorial design such as genotype x treatment

Comparison of batch correction methods in the context of DGE analysis in scRNAseq:
https://www.sciencedirect.com/science/article/pii/S200103701930409X?via%3Dihub

SVA: to deal with unknown batch effect

MetaCell - *K*-nn graph partitions

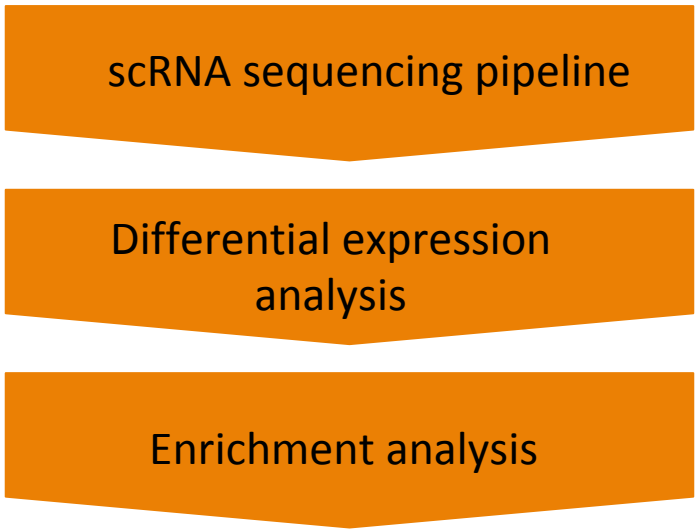




MetaCell method: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1812-2
Application to CD8 T cells: https://www.sciencedirect.com/science/article/pii/S009286741831568X

Question on DE analysis

Once we have identified DE genes, what do we do?



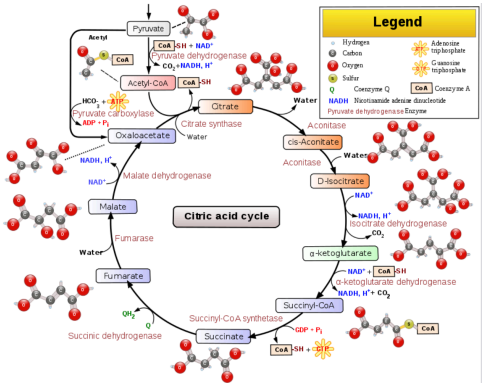
Several methods available, *e.g.*:

- over-representation analysis (ORA)
- gene set enrichment analysis (GSEA)

Goal: to gain biologically-meaningful insights from long gene lists

- test if differentially expressed genes are enriched in genes associated with a particular function
- approaches: test a small number of gene sets, or a large collection of gene sets

What is a gene set?



https://en.wikipedia.org/wiki/Citric_acid_cycle
Narayanese, WikiUserPedia, YassineMrabet, IotoBaggins – http://biocyc.org/META/NF?W-IMAGE?type=PATHWAY&object=TCA, Image adapted from Image:Citric acid cycle noLs.svg CC BY-SA 3.0

- Genes working together in a pathway (e.g. energy release through Krebs cycle)
- Genes located in the same compartment in a cell (e.g. all proteins located in the cell nucleus)
- Proteins that are all regulated by a same transcription factor
- Custom gene list that comes from a publication and that are down-regulated in a mutant
- List of genes that contain SNPs associated with a disease
- … etc!
- Several gene sets are grouped into Knowledge bases

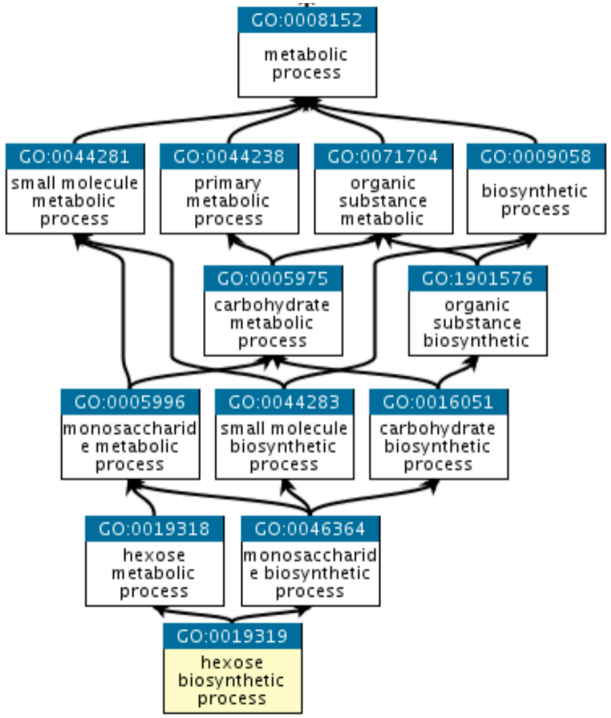
Gene ontology

- http://geneontology.org/

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.

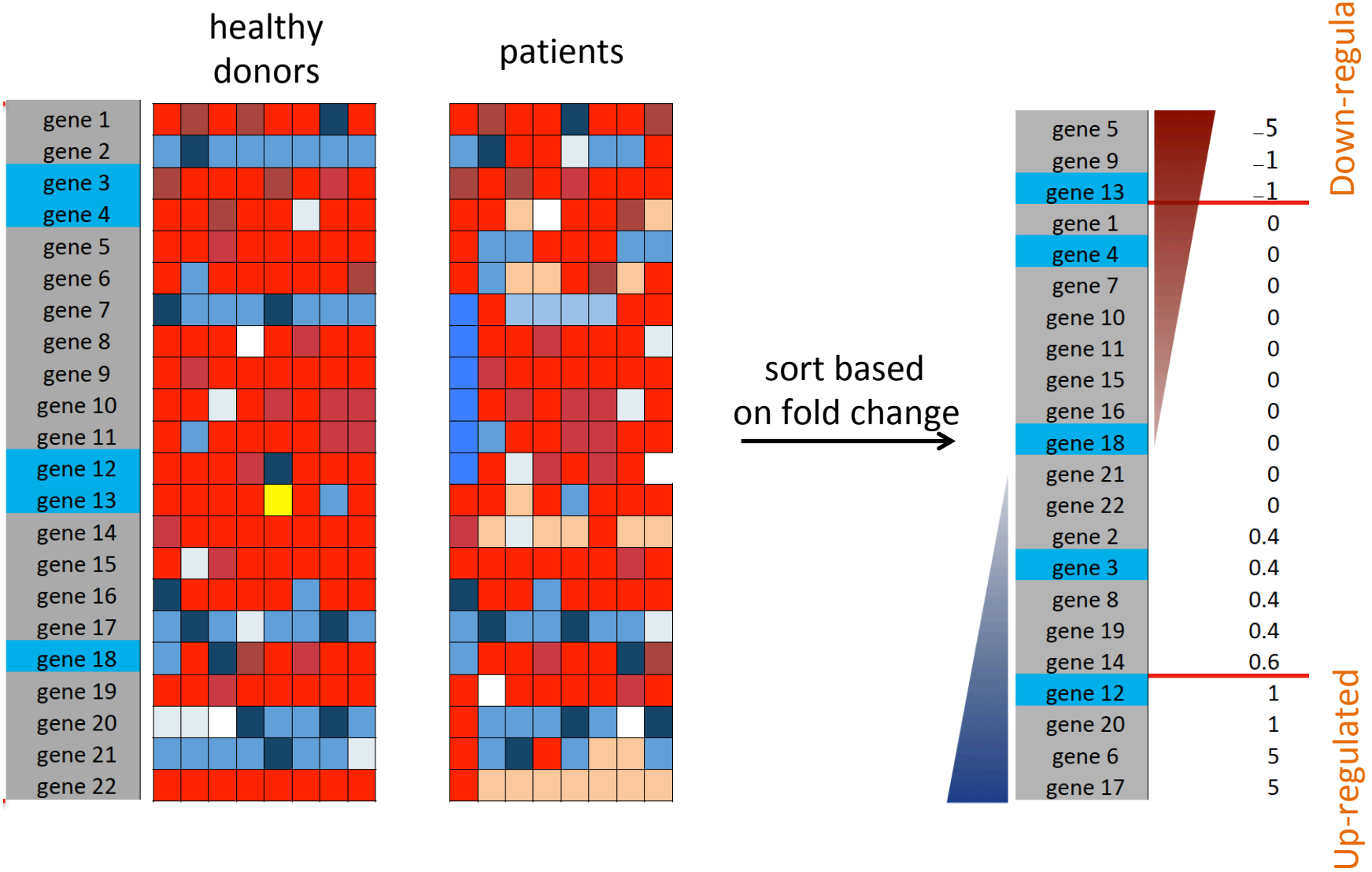
Different ontologies:
- Biological processes
- Cellular components
- Molecular functions



Sources of gene sets

- Online:
- MSigDB: database containing several types of gene set lists
- https://www.gsea-msigdb.org/gsea/msigdb/index.jsp
  - GO
  - hallmark
  - published gene sets
- KEGG (bi-directional eg mTOR signaling): https://www.kegg.jp/kegg/pathway.html
- Reactome https://reactome.org/
- WikiPathways https://www.wikipathways.org/index.php/WikiPathways

Are the genes belonging to the blue set differentially expressed?



Fisher's exact test in R

```
> cont.table<-matrix(c(2,3,5,12), ncol=2, byrow = T)
> fisher.test(cont.table)

        Fisher's Exact Test for Count Data

data:  cont.table
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  0.1012333 18.7696686
sample estimates:
odds ratio
   1.56456
```

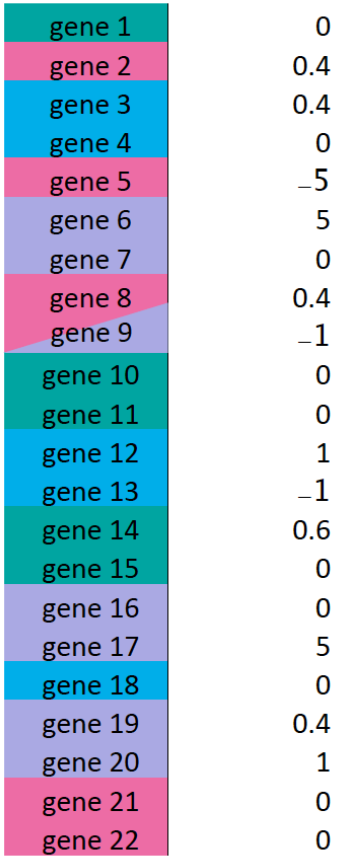
2X2 count table	Differentially expressed	Not Differentially expressed	total
blue	2	3	**5**
Not blue	5	12	**17**
total	**7**	**15**	**22**

2/5 = 0.4

5/17 = 0.3

http://mengnote.blogspot.com/2012/12/calculate-correct-hypergeometric-p.html

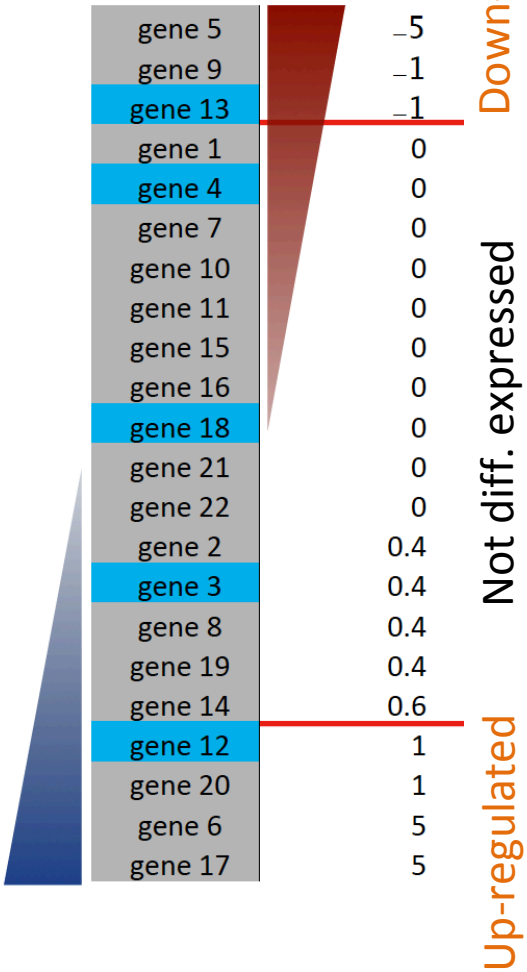
Which gene sets are differentially expressed?



Run individual Fisher's exact tests for each gene set, blue, pink, purple, green

⇒Multiple tests need p-value adjustment.

⇒Fisher test is threshold-based



Over-representation analysis using R: One possibility among many

clusterProfiler

DOI: 10.18129/B9.bioc.clusterProfiler

statistical analysis and visualization of functional profiles for genes and gene clusters

Built-in gene sets for human, mouse, yeast, etc

Built-in GO and KEGG (see later)

- ***G Yu***, LG Wang, Y Han, QY He. clusterProfiler: an R package for comparing biological themes among gene clusters. ***OMICS: A Journal of Integrative Biology*** 2012, 16(5):284-287. doi:[10.1089/omi.2011.0118](http://dx.doi.org/10.1089/omi.2011.0118)

Functions of clusterProfiler

Fisher exact test (package stats)

```
fisher.test(x, y = NULL, workspace = 200000, hybrid = FALSE,
            hybridPars = c(expect = 5, percent = 80, Emin = 1),
            control = list(), or = 1, alternative = "two.sided",
            conf.int = TRUE, conf.level = 0.95,
            simulate.p.value = FALSE, B = 2000)
```

Over-representation analysis (similar to Fisher test) for built-in GO gene sets:

```
enrichGO(gene, OrgDb, keyType = "ENTREZID", ont = "MF",
  pvalueCutoff = 0.05, pAdjustMethod = "BH", universe,
  qvalueCutoff = 0.2, minGSSize = 10, maxGSSize = 500,
  readable = FALSE, pool = FALSE)
```

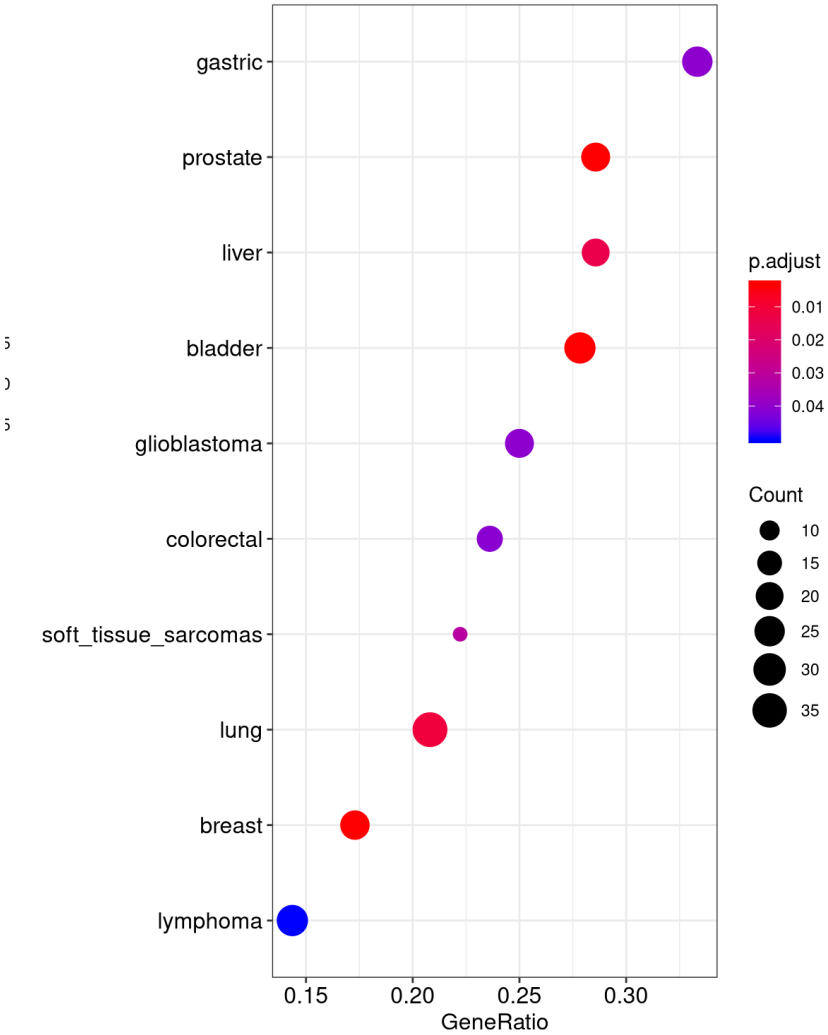
enricher(): similar enrichGO() but for user defined gene sets

```
enricher(gene, pvalueCutoff = 0.05, pAdjustMethod = "BH", universe,
  minGSSize = 10, maxGSSize = 500, qvalueCutoff = 0.2, TERM2GENE,
  TERM2NAME = NA)
```

Visualizations available in clusterProfiler

- dotplot

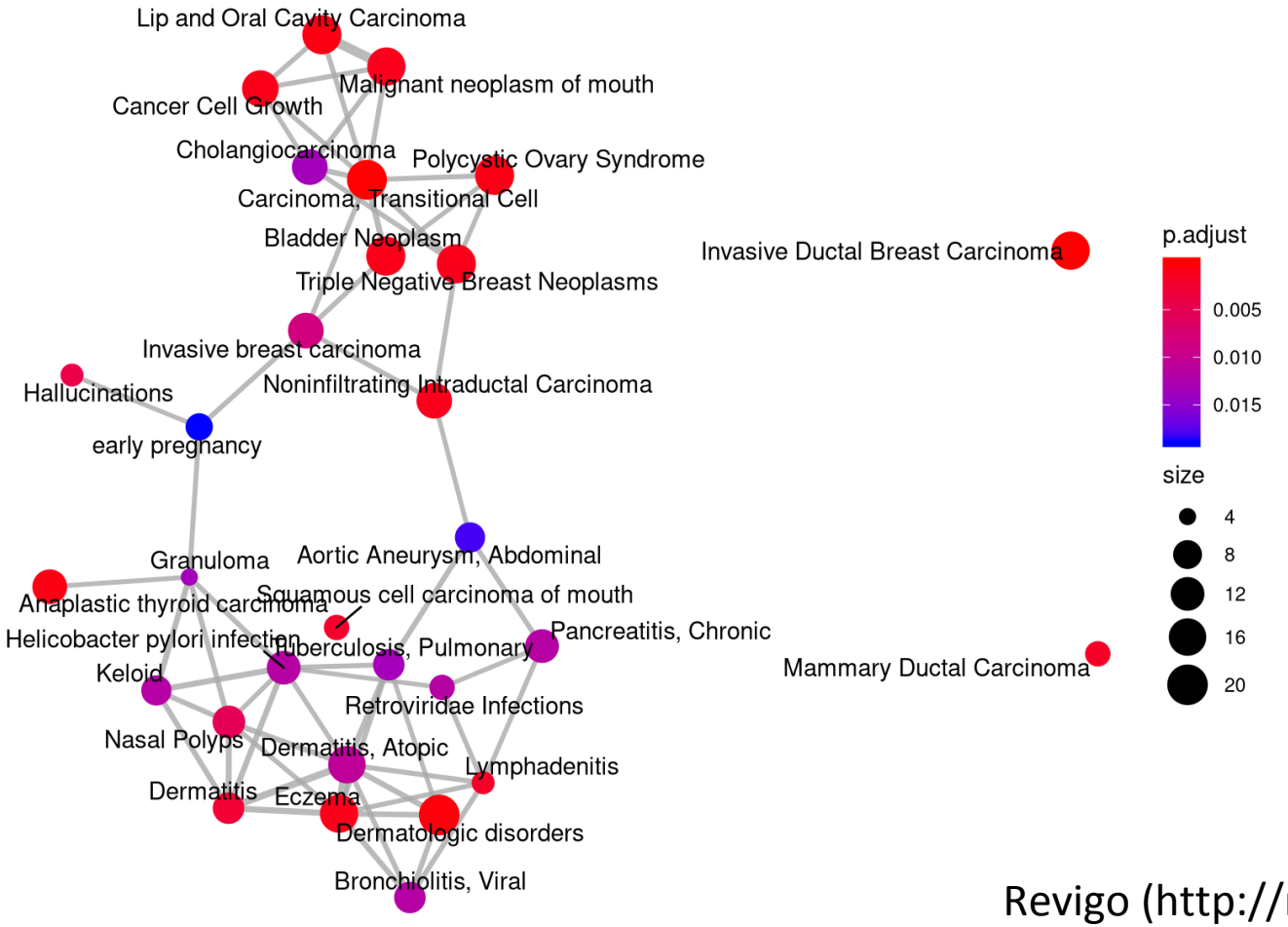
```
ego <- enrichGO(de)
dotplot(ego, showCategory=10)
```



Visualizations available in clusterProfiler

- Enrichment map

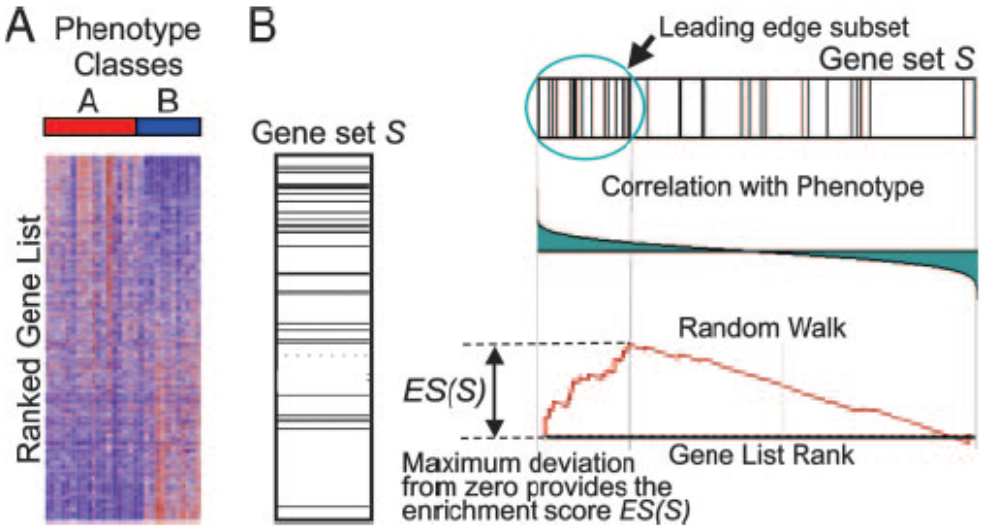
```
ego <- enrichGO(de)
emapplot(pairwise_termsim(ego))
```



Revigo (http://revigo.irb.hr/)

Gene set enrichment analysis (GSEA)

Can be performed if you have statistics for all genes detected in the scRNAseq dataset, when using limma or edgeR



gseGO(): GSEA of GO gene sets using all ranked genes

```
gseGO(geneList, ont = "BP", OrgDb, keyType = "ENTREZID",
  exponent = 1, nPerm = 1000, minGSSize = 10, maxGSSize = 500,
  pvalueCutoff = 0.05, pAdjustMethod = "BH", verbose = TRUE,
  seed = FALSE, by = "fgsea")
```

Question on enrichment analysis