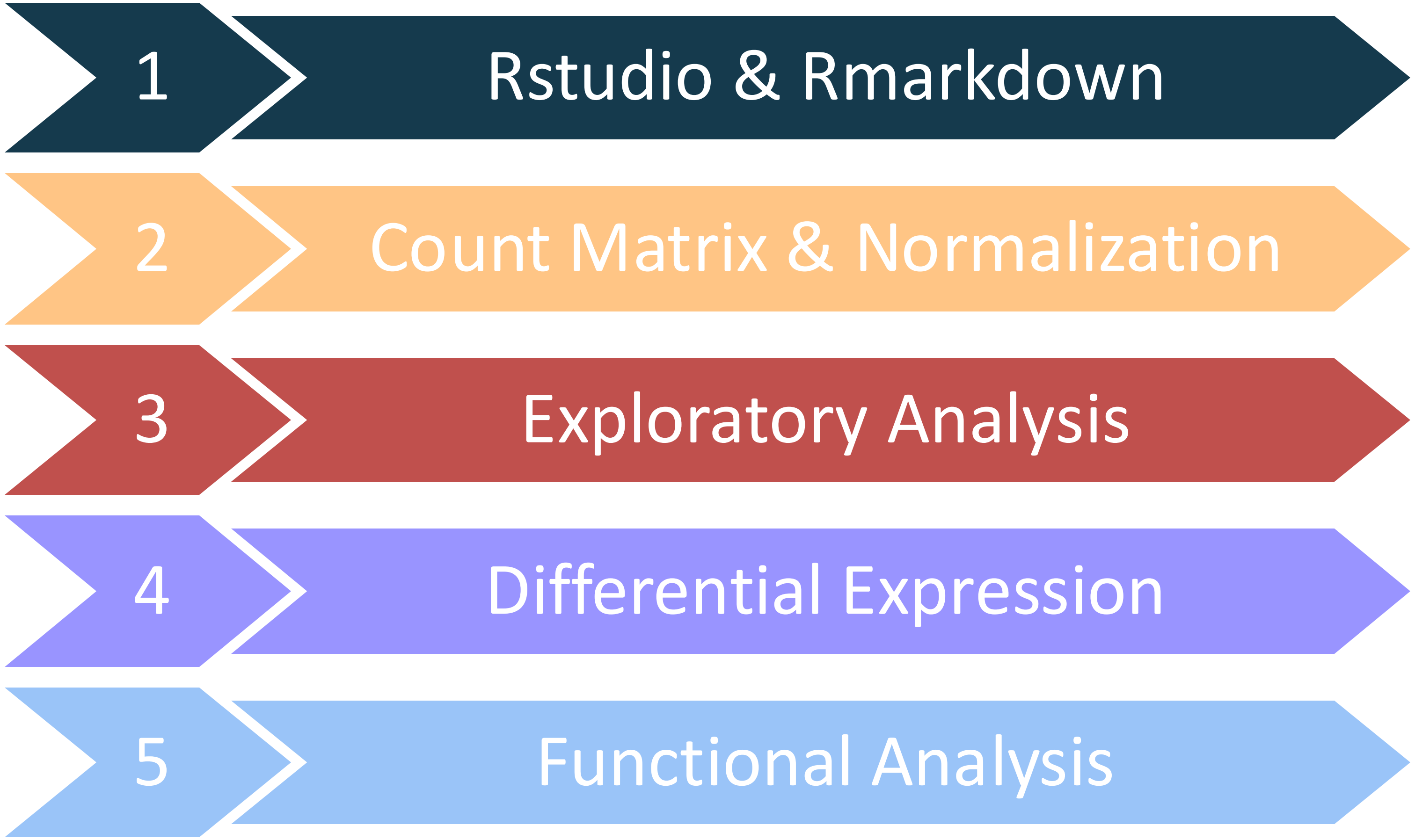
Functional Analysis

Center for Health Data Science

Health Data Science Sandbox

Overview



1 Rstudio & Rmarkdown

2 Count Matrix & Normalization

3 Exploratory Analysis

4 Differential Expression

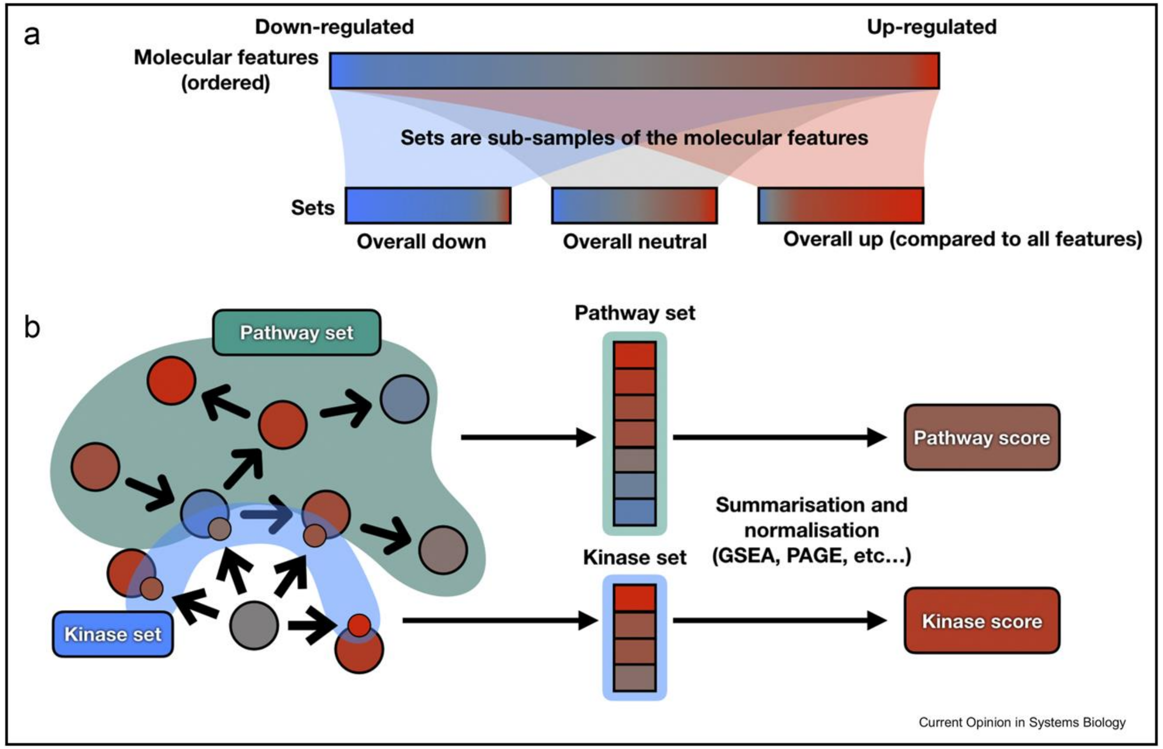
5 Functional Analysis

HeaDS

Enrichment analysis

Enrichment Analysis (EA):

- Identify groups of genes that are over-represented within a larger set of genes
- Enriched sets of genes may be associated with biological pathways and processes
- Returns scores/ranks and p-values
- Some types are:
  - SEA (Singular EA)
  - GSEA (Gene Set EA)
  - MEA (Modular EA)


a
Down-regulated
Up-regulated
Molecular features (ordered)
Sets are sub-samples of the molecular features
Sets
Overall down
Overall neutral
Overall up (compared to all features)
b
Pathway set
Pathway set
Pathway score
Summarisation and normalisation (GSEA, PAGE, etc...)
Kinase set
Kinase set
Kinase score
Current Opinion in Systems Biology

https://doi.org/10.1016/j.coisb.2019.04.002

Enrichment analysis

- Are my differentially expressed genes enriched for Kinases-related ontology term(s)?
- Create a contingency table where:

	Kinase-related otology	Other ontology
DE	30	120
Not DE	40	960

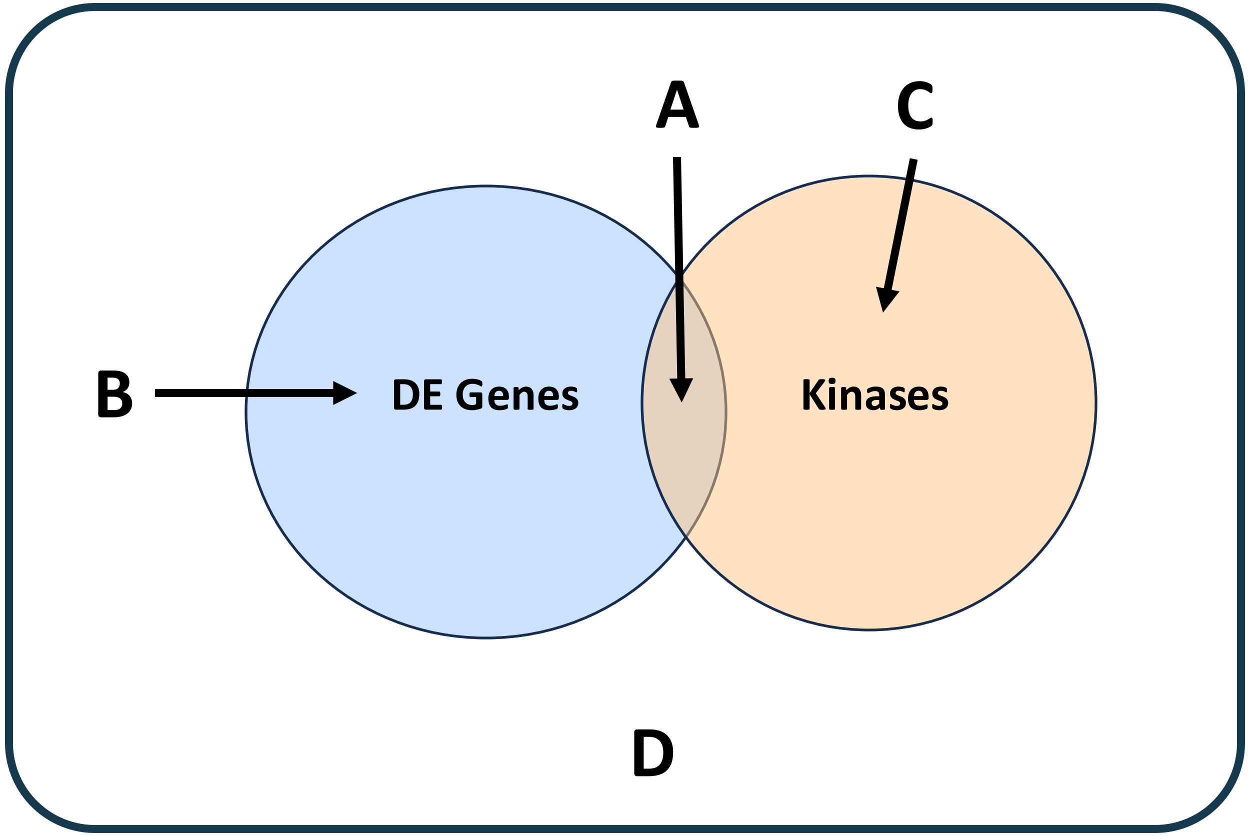
- Perform a Fisher's exact test to check if there is enrichment → **p-value**
- There are thousands of GO sets, so multiple testing correction is needed

Enrichment analysis

Is the overlap (**A**) of these two gene sets higher than what we would expect if they were independent?

	Kinase	Not kinase
DE	A	B
Not DE	C	D

All genes

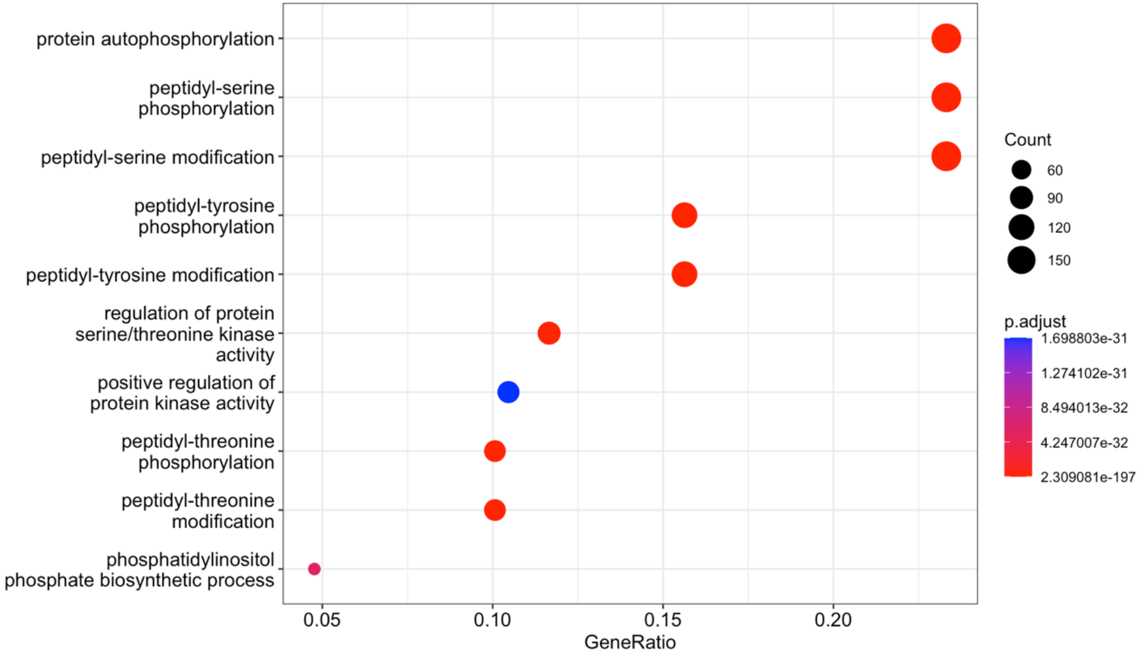


Enrichment analysis

Quiz: Enrichment Analysis on a kinase screen

- Our experiment returns a list of kinases which were regulated.
- We perform EA on this list and use the whole proteome (transcriptome) as the background.
- What do you expect to see enriched?

	Kinase	Not kinase
Regulated	A	B
Not Regulated	C	D

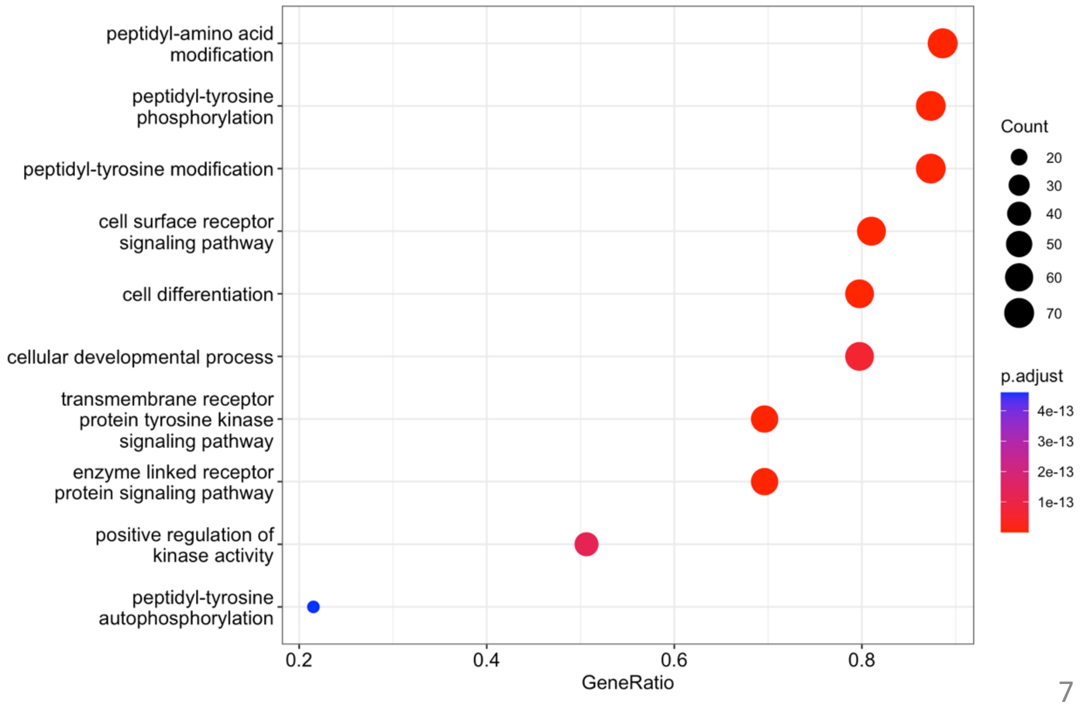


Enrichment analysis

Quiz: Enrichment Analysis on a kinase screen

- Change the gene background!
- Instead of all genes, **use a custom background**: only kinase-ome
- What do we achieve by doing this?

	Kinase subclass	Other kinases
Gene list	A	B
Not gene list	C	D



Enrichment analysis

You investigate differential gene expression in **mouse liver tissue response to a kinase inhibitor.** You have obtained a list of DEGs.
What is the most appropriate background to find overrepresented GO terms involved in the drug response?

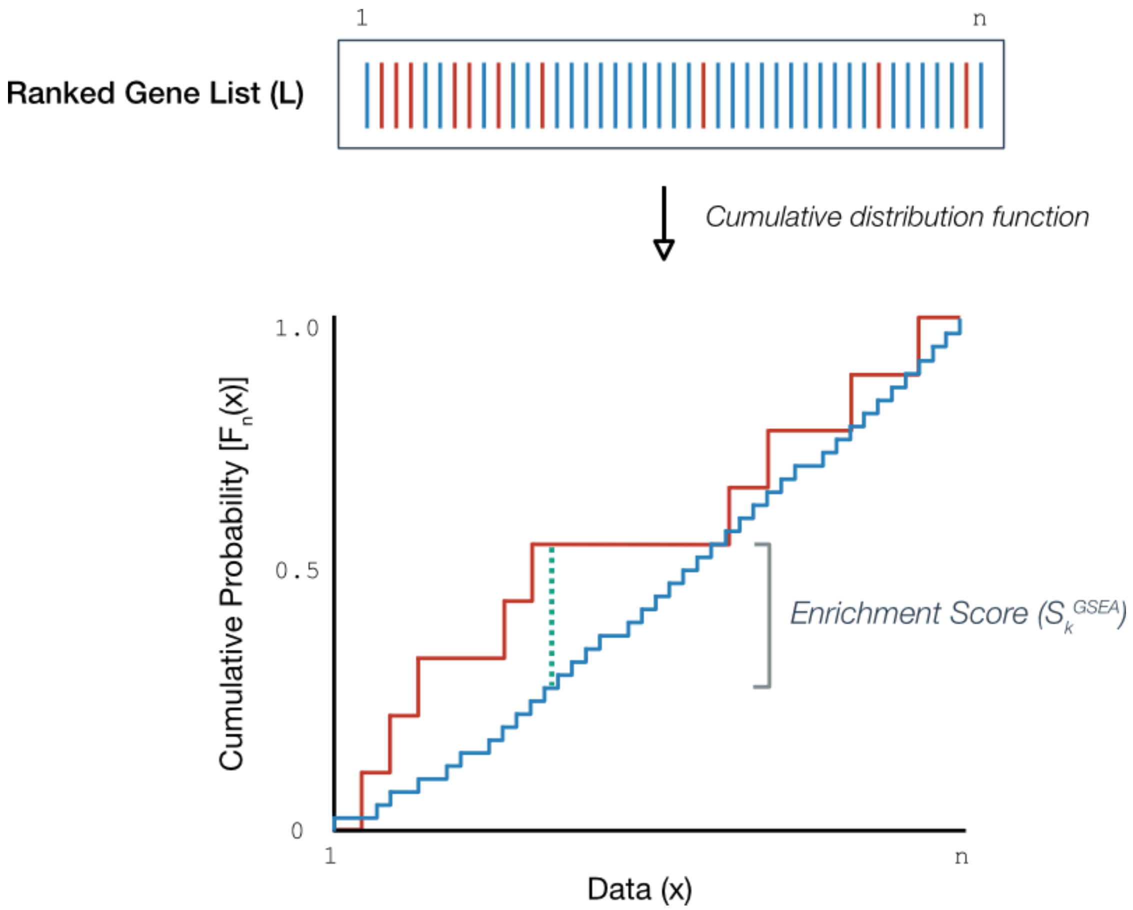
A. All mouse proteins (genes) that are phosphorylated
B. Genes (proteins) in mouse liver kinase-ome
C. Genes expressed in your experiment

How would you produce this list?

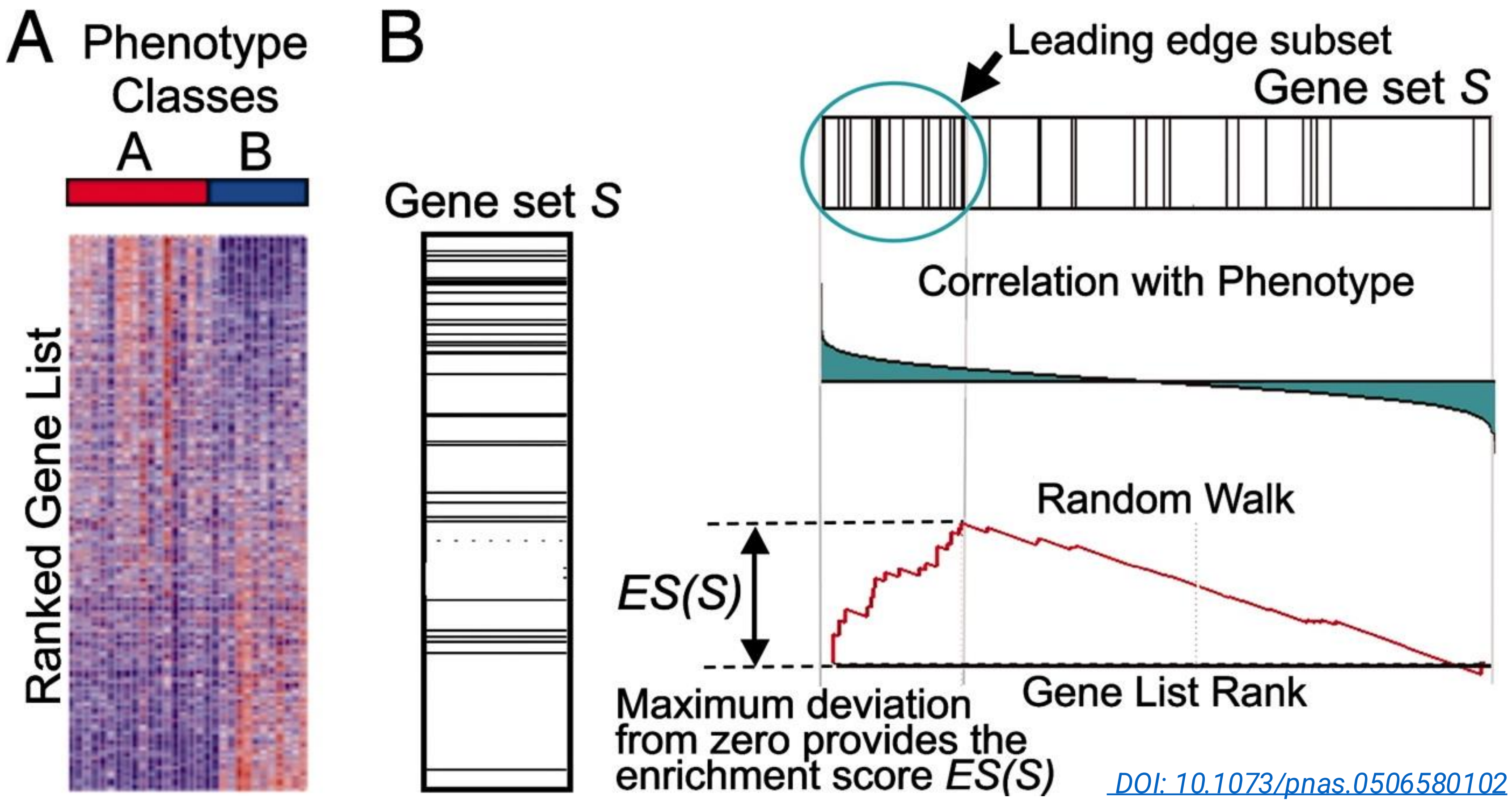
GSEA

Example: **GSEA or class scoring**

- Are my DE genes enriched for Kinases (Gene Ontology)?
- Rank my DE results by log2FC
- Running enrichment score
- Test significance of max enrich. score



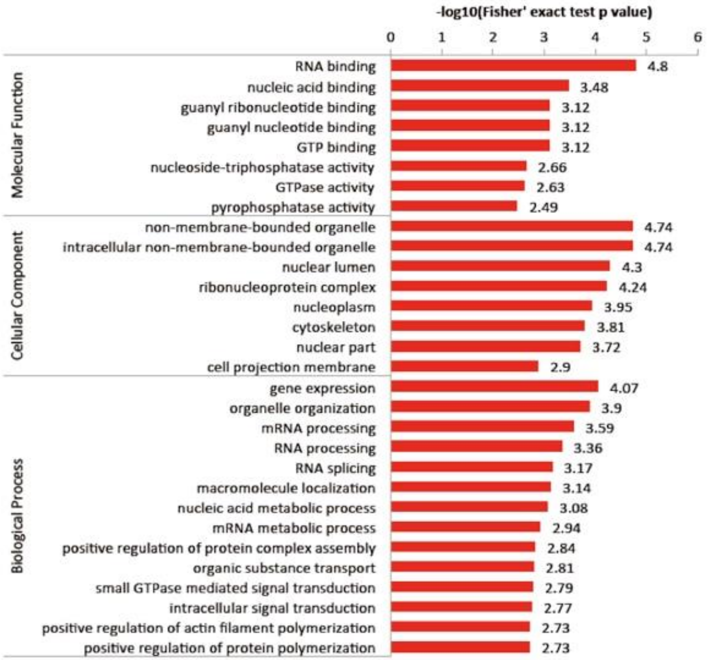
GSEA

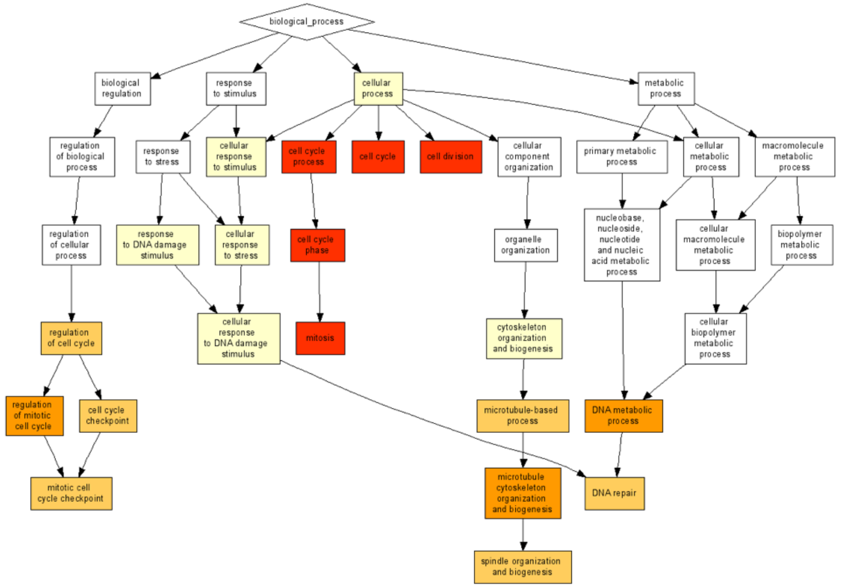


DOI: 10.1073/pnas.0506580102

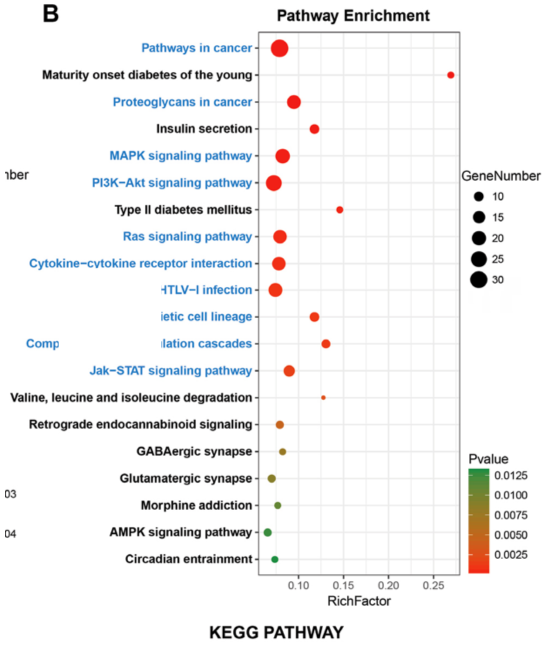
Enrichment analysis Tools

- **Web Resources:** KEGG, PANTHER, DAVID, DO
- **R packages:** clusterProfiler, GOSemSim



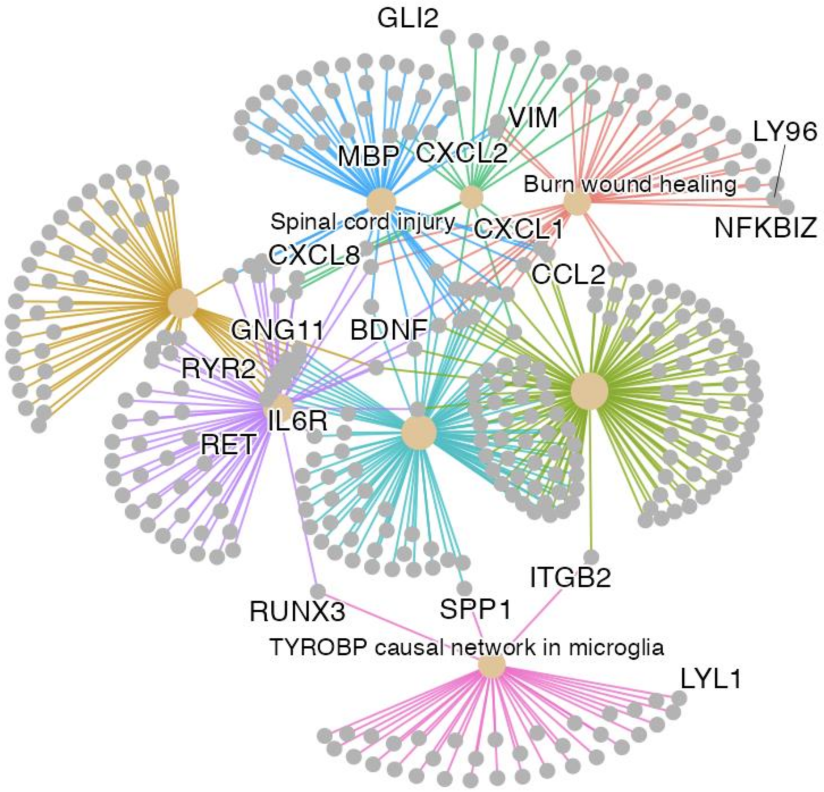


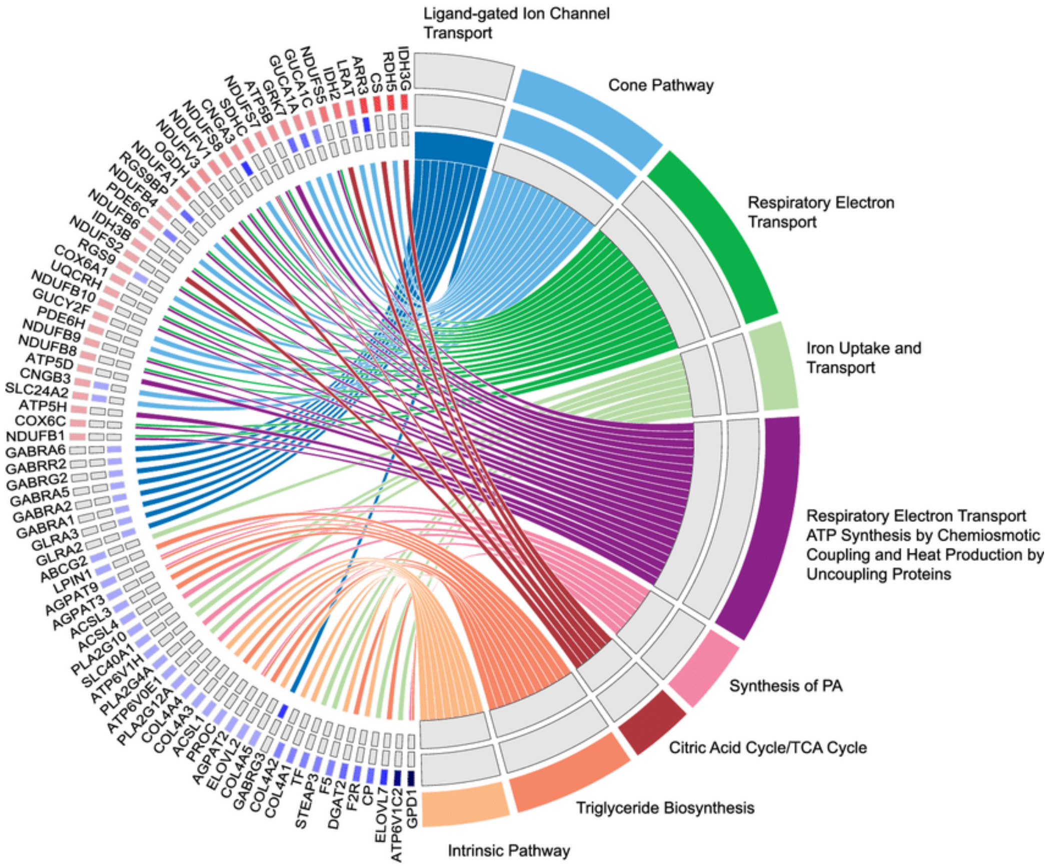
biological_process
biological regulation
response to stimulus
cellular process
metabolic process
regulation of biological process
response to stress
cellular response to stimulus
cell cycle process
cell cycle
cell division
cellular component organization
primary metabolic process
cellular metabolic process
macromolecule metabolic process
regulation of cellular process
response to DNA damage stimulus
cellular response to stress
cell cycle phase
organelle organization
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
cellular macromolecule metabolic process
biopolymer metabolic process
regulation of cell cycle
cellular response to DNA damage stimulus
mitosis
cytoskeleton organization and biogenesis
cellular biopolymer metabolic process
regulation of mitotic cell cycle
cell cycle checkpoint
microtubule-based process
DNA metabolic process
mitotic cell cycle checkpoint
microtubule cytoskeleton organization and biogenesis
DNA repair
spindle organization and biogenesis



CUSTOM PLOTS

- **R packages:** igraph, visNetwork, ggnetwork, ggnet, circlize



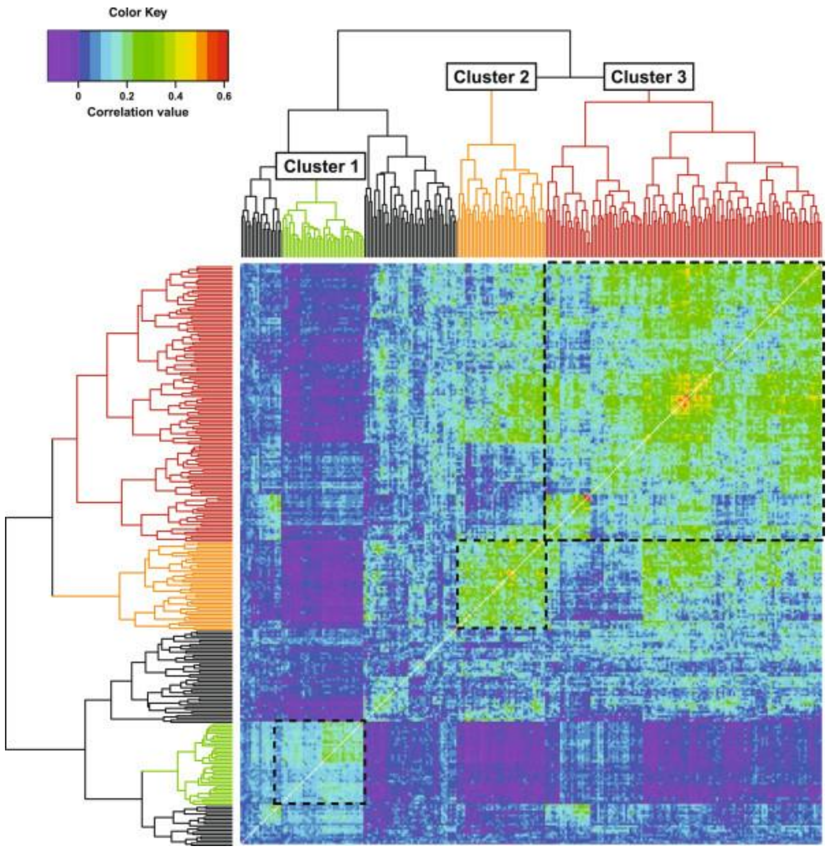


HeaDS

https://www.nature.com/articles/s41598-021-84338-y

Co-expression Analysis

- Identify clusters of correlated genes, based on expression across samples within a condition
- Couple these to clinical variables and patient metadata
- Co-expression clusters can be used for enrichment -and pathway analysis.
- Tools:
  - DGCA (Differential Gene Correlation Analysis) https://doi.org/10.1186/s12918-016-0349-1
  - WGCNA (Weighted Gene Co-expression network Analysis) https://doi.org/10.1186/1471-2105-9-559



Networks Analysis

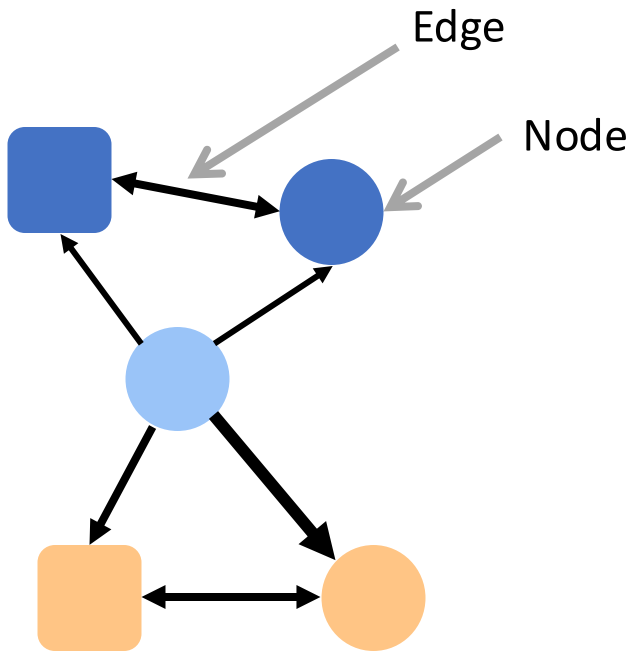
Genes of interest can be used to construct networks

Network: set of interconnected nodes

- Nodes: items we want to connect (genes, proteins, etc)
- Edges: relationship between nodes (correlation, score)

Attributes: Information about the network

- Color and shape: e.g. differentiate genes from diseases
- Arrow head: direction of the relationship
- Edge thickness: strength of relationship

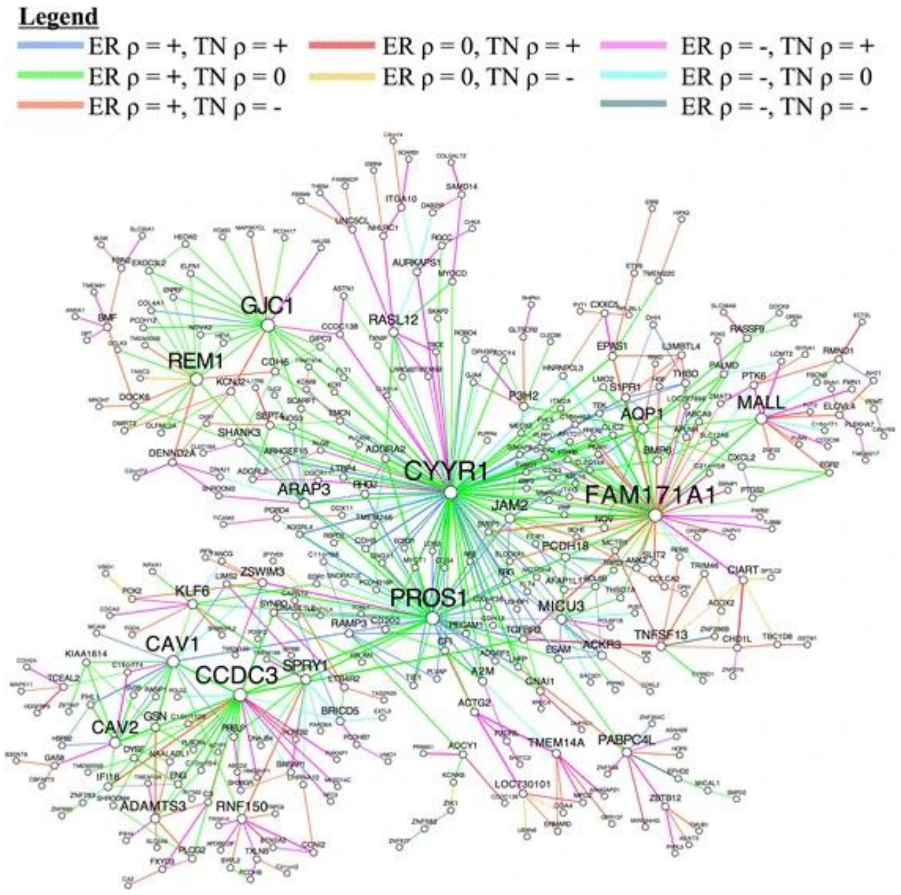


Interaction Networks

- Gene correlations for differential changes between conditions

- Annotation of the network:
  - diseases or drugs (Clinical Knowledge Graphs)
  - protein interactions
  - transcription factors, etc.

- Tool interaction networks:
  - Mining Interactions: STRING https://string-db.org/
  - Network Visualization: Cytoscape https://cytoscape.org/



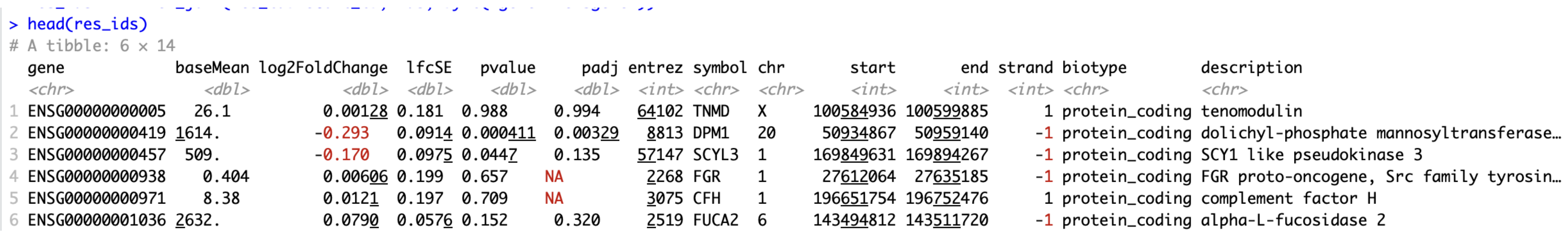
Exercise

Let’s do some Functional Analysis!

- Notebooks:
  - *08b_FA_overrepresentation.Rmd*
  - *08c_FA_GSEA.Rmd*

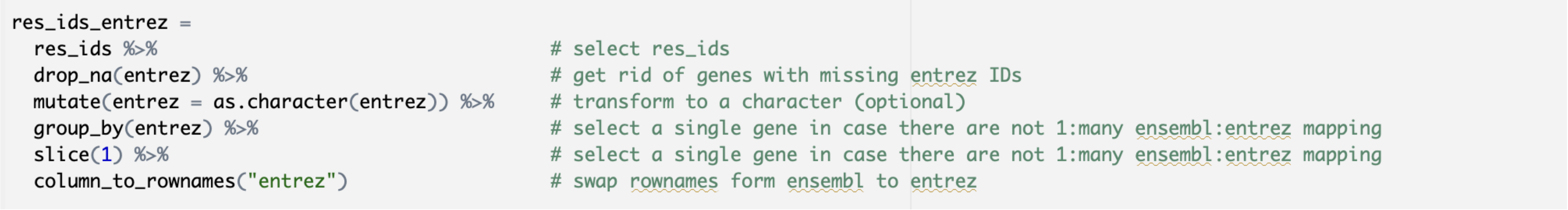
How to convert gene IDs

problem: genes have ensembl IDs (ENSG…) but we need entrez IDs



```
> head(res_ids)
# A tibble: 6 × 14
  gene            baseMean log2FoldChange  lfcSE   pvalue     padj entrez symbol chr       start       end strand biotype        description
  <chr>              <dbl>          <dbl>  <dbl>    <dbl>    <dbl>  <int> <chr>  <chr>     <int>     <int>  <int> <chr>          <chr>
1 ENSG00000000005   26.1          0.00128 0.181  0.988     0.994    64102 TNMD   X     100584936 100599885      1 protein_coding tenomodulin
2 ENSG00000000419 1614.          -0.293   0.0914 0.000411  0.00329   8813 DPM1   20     50934867  50959140     -1 protein_coding dolichyl-phosphate mannosyltransferase…
3 ENSG00000000457  509.          -0.170   0.0975 0.0447    0.135    57147 SCYL3  1     169849631 169894267     -1 protein_coding SCY1 like pseudokinase 3
4 ENSG00000000938    0.404        0.00606 0.199  0.657    NA         2268 FGR    1      27612064  27635185     -1 protein_coding FGR proto-oncogene, Src family tyrosin…
5 ENSG00000000971    8.38         0.0121  0.197  0.709    NA         3075 CFH    1     196651754 196752476      1 protein_coding complement factor H
6 ENSG00000001036 2632.           0.0790  0.0576 0.152     0.320     2519 FUCA2  6     143494812 143511720     -1 protein_coding alpha-L-fucosidase 2
```

Solution1: swap rownames from ensembl to entrez



```
res_ids_entrez =
  res_ids %>%                                   # select res_ids
  drop_na(entrez) %>%                           # get rid of genes with missing entrez IDs
  mutate(entrez = as.character(entrez)) %>%     # transform to a character (optional)
  group_by(entrez) %>%                          # select a single gene in case there are not 1:many ensembl:entrez mapping
  slice(1) %>%                                  # select a single gene in case there are not 1:many ensembl:entrez mapping
  column_to_rownames("entrez")                  # swap rownames form ensembl to entrez
```

HeaDS