



EpiSyStem: Bioinformatics Workshop
Istanbul – 2018

Ege Ulgen, MD

Brief Introduction

pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks

 Ege Ulgen, Ozan Ozisik,  Osman Ugur Sezerman

doi: <https://doi.org/10.1101/272450>

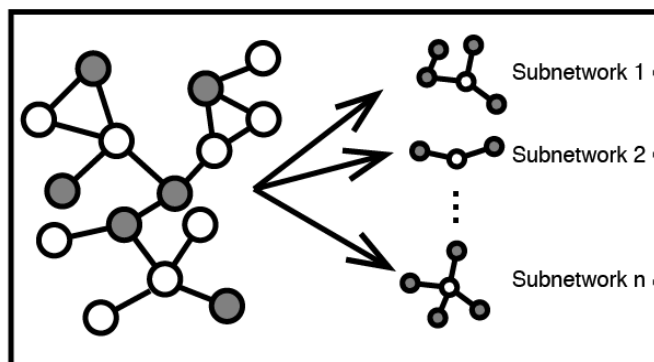
- Using input genes, pathfindR identifies sets of genes that form **active subnetworks** in a protein-protein interaction network
- It then performs **pathway enrichment analyses** on the identified gene sets
- Additionally, **clusters** these pathways and identify representative pathways in the clusters

Enrichment Workflow

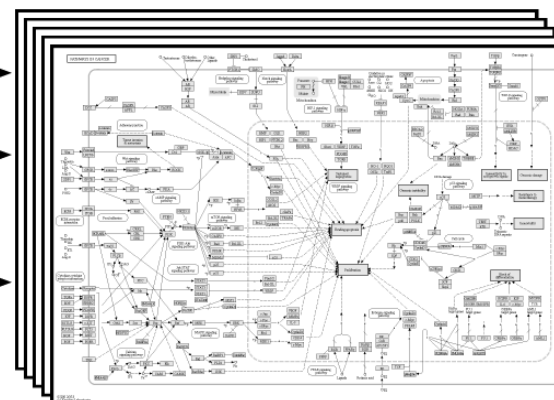
**Input
Data Frame**

Gene	LFC	p-value
...

Active Subnetwork Search



**Pathway Enrichment
Analyses**

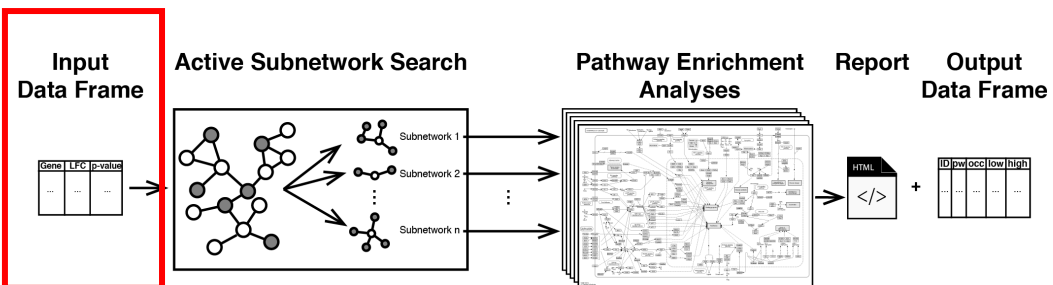


Report

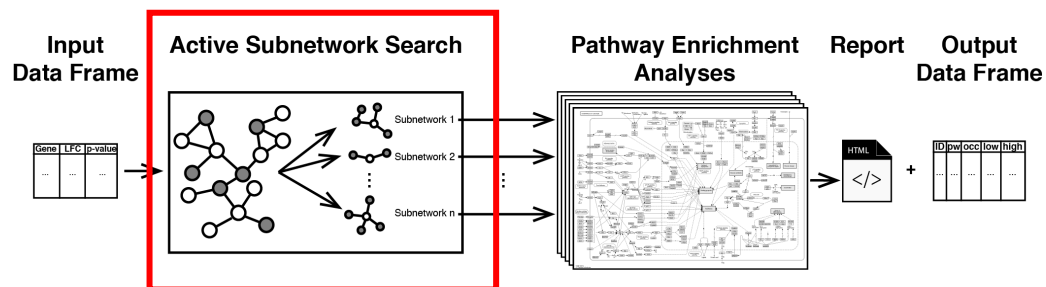


**Output
Data Frame**

ID	pw	occ	low	high
...



Gene Symbol	Change Value (OPTIONAL)	Adjusted p-value
FAM110A	-0.6939359	0.0000034
RNASE2	1.3535040	0.0000101
S100A8	1.5448338	0.0000347
S100A9	1.0280904	0.0002263
TEX261	-0.3235994	0.0002263
ARHGAP17	-0.6919330	0.0002708

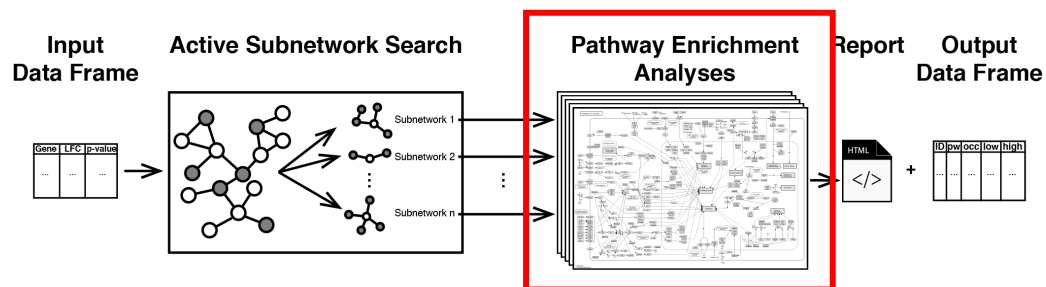


• Active Subnetwork Search Algorithms:

- Greedy Algorithm
- Simulated Annealing
- Genetic Algorithm

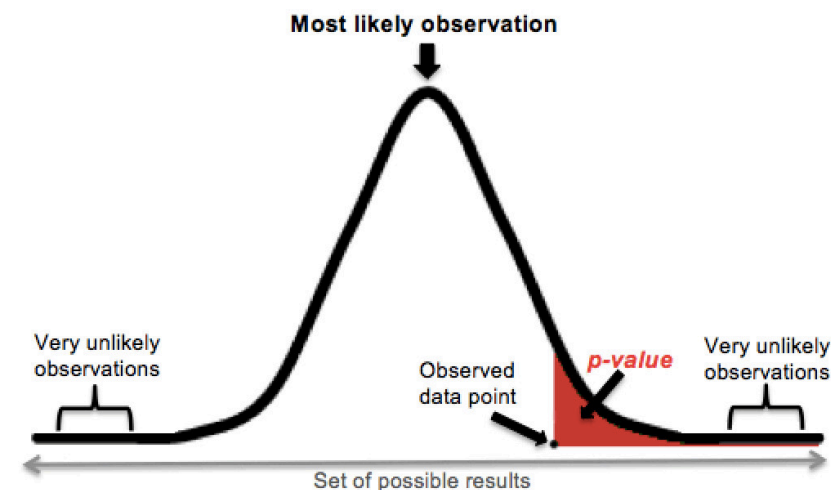
• Available Protein Interaction Networks (PINs):

- Biogrid
- GeneMania
- IntAct
- KEGG PIN
- Custom PIN

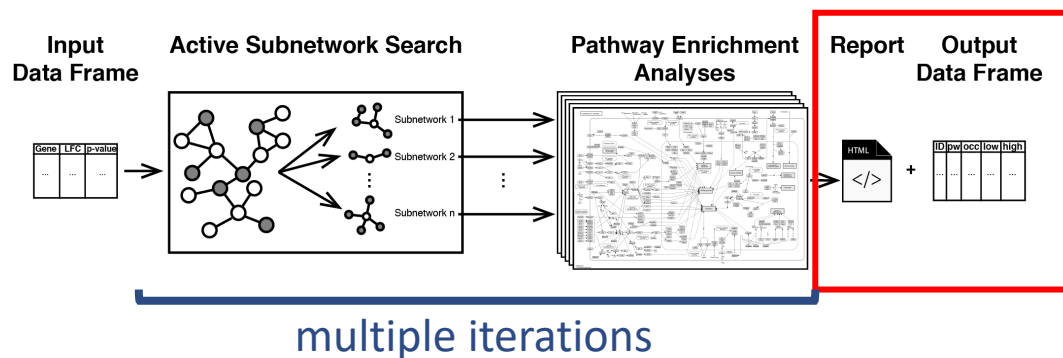


- Available gene sets:
 - KEGG
 - Reactome
 - BioCarta
 - Gene Ontology gene sets
 - GO – All
 - GO – BP
 - GO – CC
 - GO – MF
 - Custom gene sets

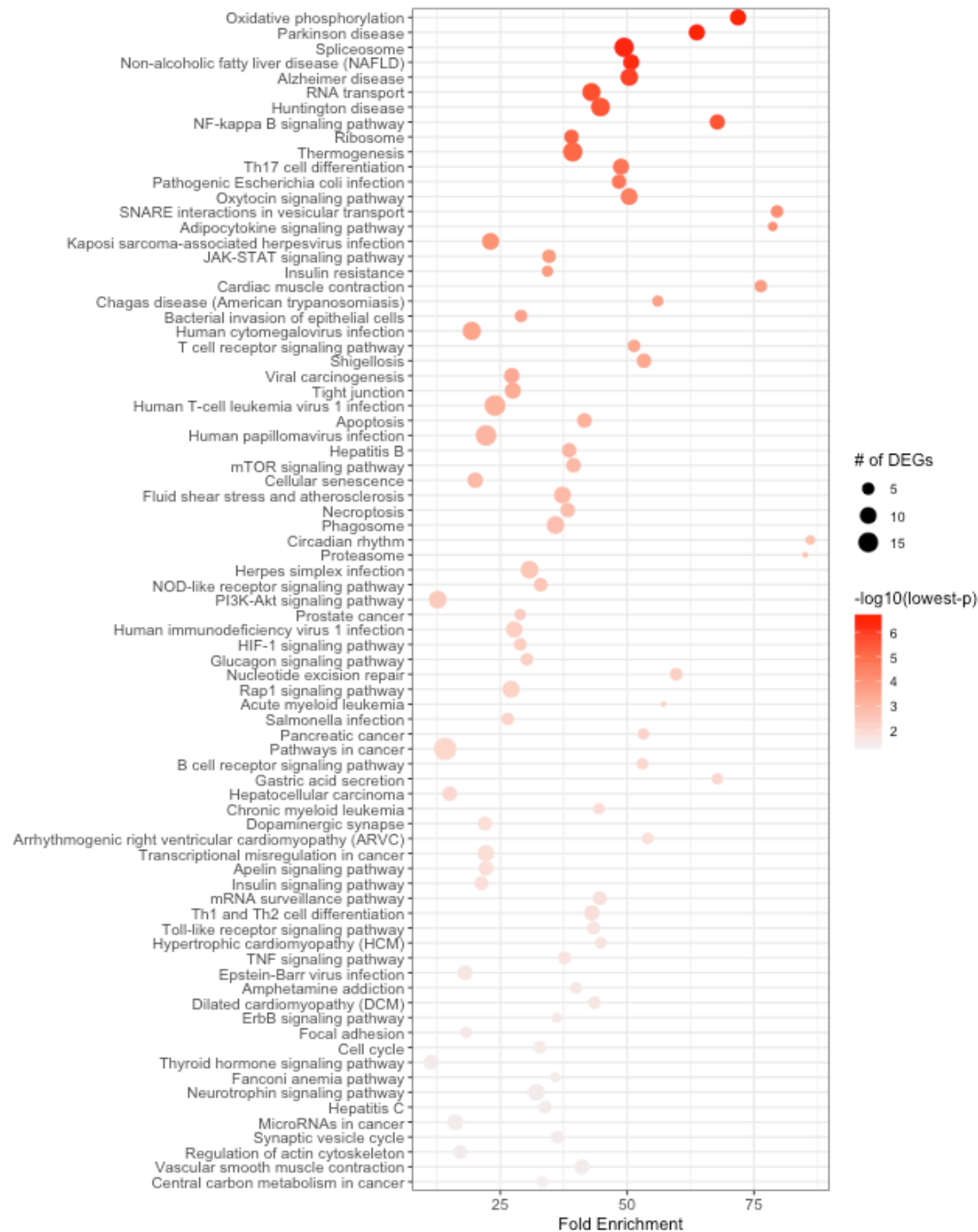
$$P(X = k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$$

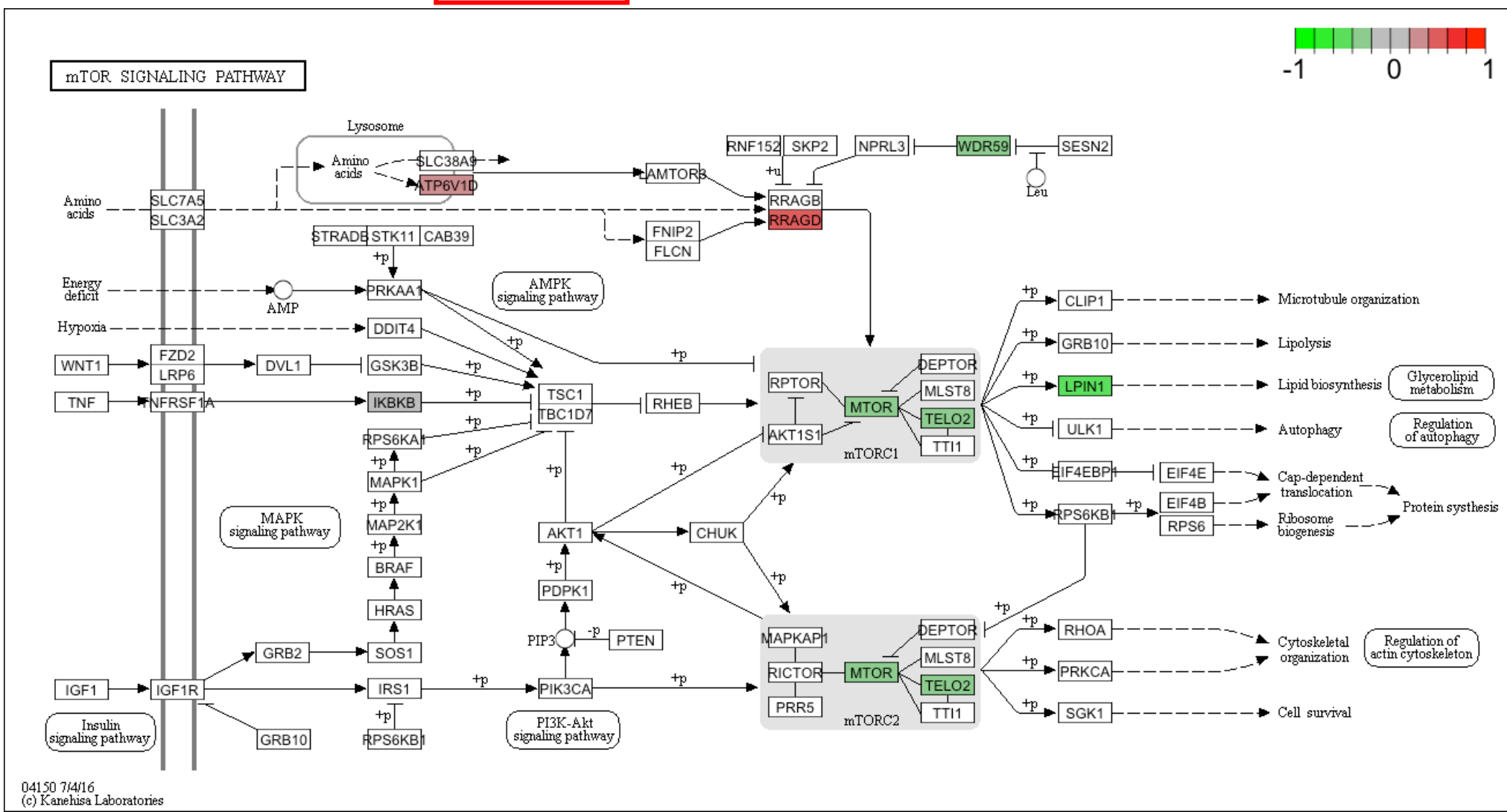
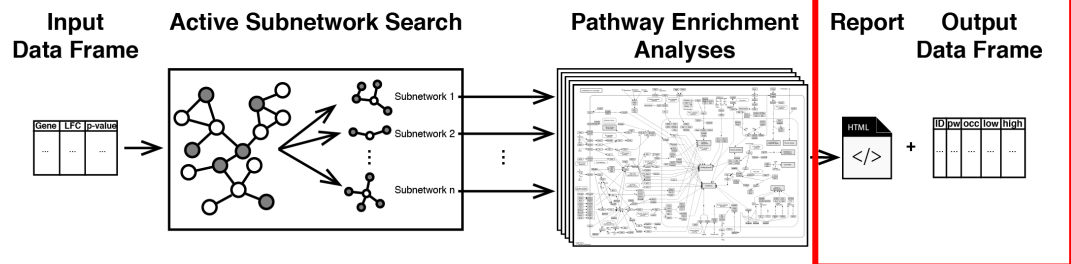


A **p-value** (shaded red area) is the probability of an observed (or more extreme) result arising by chance



ID	Pathway	Fold_Enrichment	occurrence	lowest_p	highest_p	Up_regulated	Down_regulated
hsa00190	Oxidative phosphorylation	71.86252	10	3e-07	3e-07	NDUFB3, NDUFA1, COX7C, COX7A2, UQCRQ, COX6A1, ATP6V0E1, ATP6V1D	ATP6V0E2
hsa05012	Parkinson's disease	63.72714	10	4e-07	4e-07	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	SLC25A5, VDAC1, UBE2G1



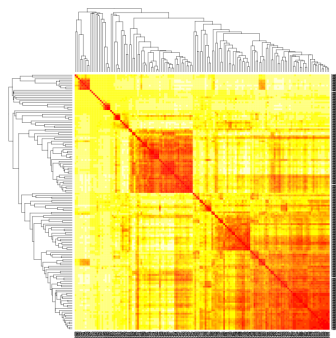


Clustering Workflow

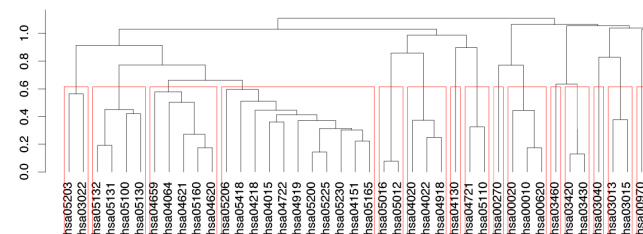
Input Pathways

ID	pw	occ	low	high
...

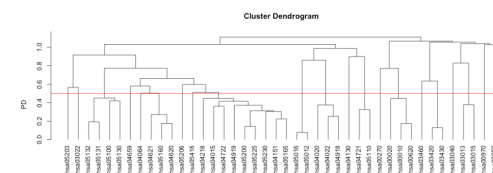
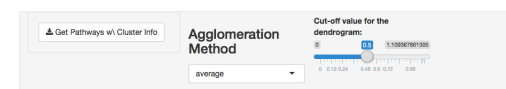
Calculation of Pairwise Distances



Hierarchical Clustering & Partitioning into Clusters



automatic

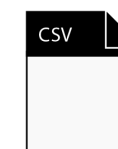


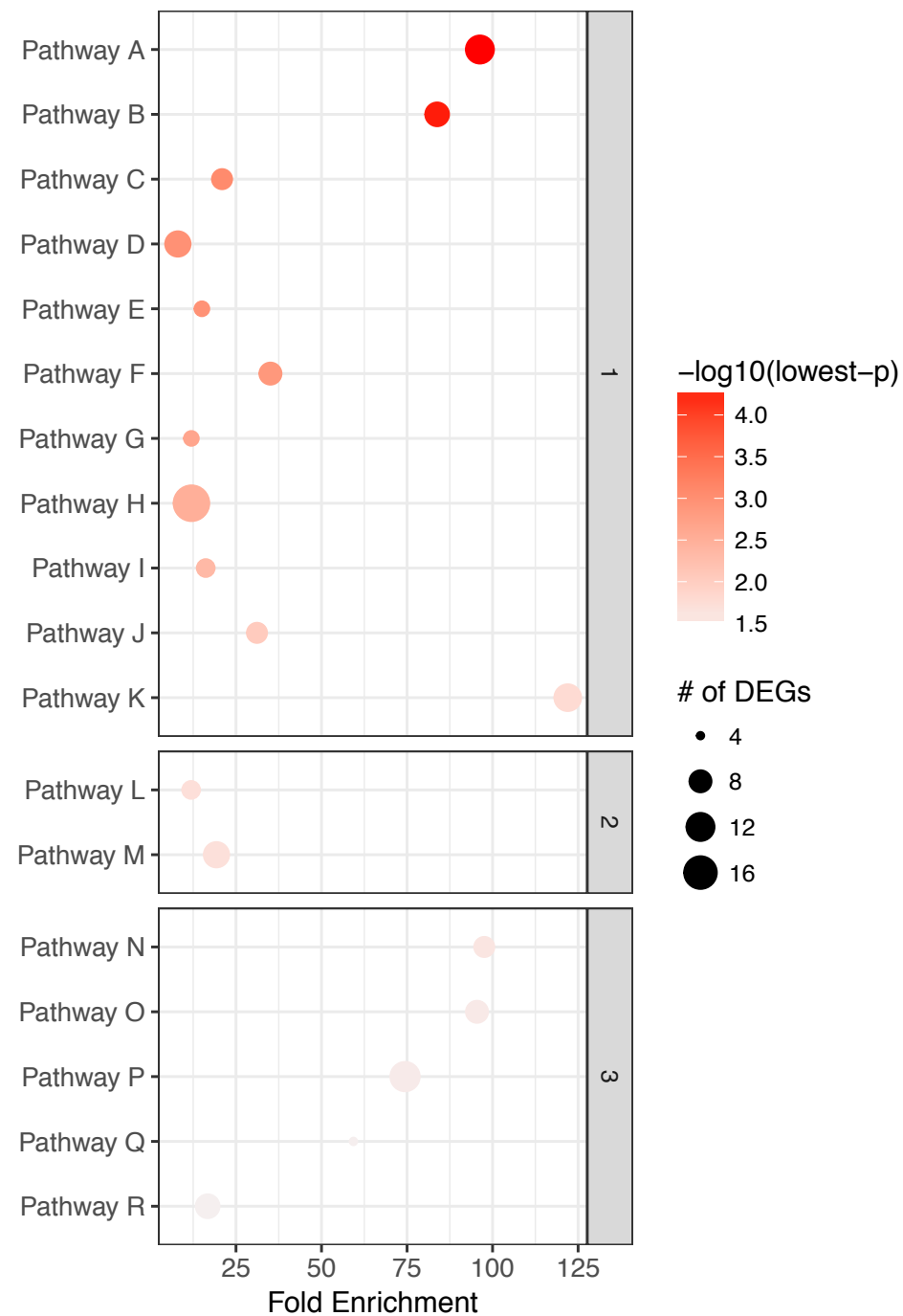
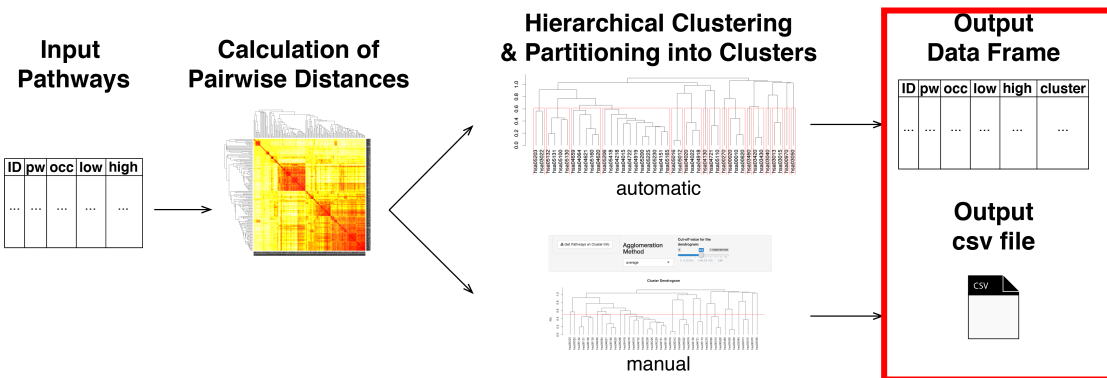
manual

Output Data Frame

ID	pw	occ	low	high	cluster
...

Output csv file





Pathway Scoring

For a set of pathways $P = \{P_1, P_2, \dots, P_n\}$, where each P_i contains a set of genes, i.e. $P_i = \{g_1, g_2, \dots, g_k\}$, the pathway score matrix PS is defined as:

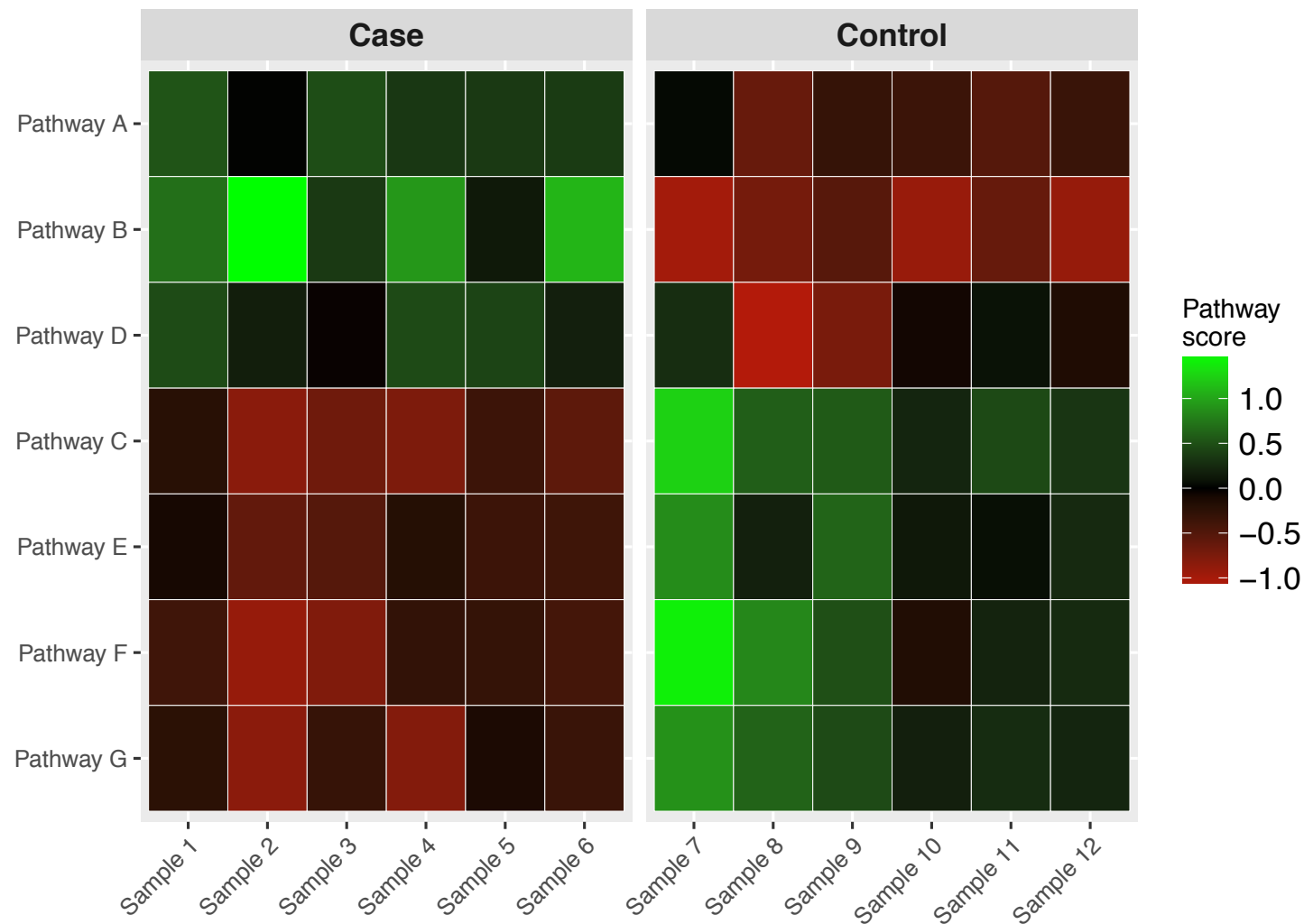
$$2 \text{ } PS_{p,s} = \frac{1}{k} \sum_{g \in P_p} GS_{g,s} \text{ for each pathway } p \text{ and for each sample } s.$$

GS is the gene score per sample matrix and is defined as:

$$1 \text{ } GS_{g,s} = (EM_{g,s} - \bar{x}_g) / s_g$$

where EM is the expression matrix (columns are samples, rows are genes), \bar{x}_g is the mean expression value of the gene and s_g is the standard deviation of the expression values for the gene.

Pathway Scoring



Demonstration



Demo – I – Installation

Installation – Dependencies

```
if (!requireNamespace("BiocManager", quietly = TRUE))  
  install.packages("BiocManager")  
BiocManager::install("pathview", version = "3.8")  
BiocManager::install("AnnotationDbi", version = "3.8")  
BiocManager::install("org.Hs.eg.db", version = "3.8")
```

Installation – pathfindR

```
install.packages("pathfindR")
```



Demo – II – Pathway Enrichment

```
library(pathfindR)
```

```
RA_demo <- run_pathfindR(RA_input,  
                          gene_sets = "BioCarta",  
                          pin_name_path = "GeneMania",  
                          output = "DEMO_OUTPUT")
```

Demo – III – Pathway Clustering

```
RA_clustered <- choose_clusters(RA_demo)
```

OR

```
choose_clusters(RA_demo, auto = FALSE)
```

Demo – IV – Pathway Scoring

selecting "Representative" pathways for clear visualization

```
pws_table <- RA_clustered[RA_clustered$Status == "Representative", ]
```

Expression matrix

```
exp_mat <- pathfindR::RA_exp_mat
```

Vector of "Case" IDs

```
cases <- c("GSM389703", "GSM389704", "GSM389706",  
"GSM389708", "GSM389711", "GSM389714", "GSM389716",  
"GSM389717", "GSM389719", "GSM389721", "GSM389722",  
"GSM389724", "GSM389726", "GSM389727", "GSM389730",  
"GSM389731", "GSM389733", "GSM389735")
```

Calculate pathway scores and plot heatmap

```
score_matrix <- calculate_pw_scores(pws_table, exp_mat, cases)
```



Resources

- Tutorial on Biostars:
 - <https://www.biostars.org/p/322415/>
- Vignette
 - https://cran.r-project.org/web/packages/pathfindR/vignettes/pathfindr_vignette.html
- pathfindR Wiki:
 - <https://github.com/egeulgen/pathfindR/wiki>
- To report any issues:
 - <https://github.com/egeulgen/pathfindR/issues>
- For all other questions:
 - egeulgen@gmail.com