pathfindR

# COST CHARME SUMMER TRAINING SCHOOL

ISTANBUL – 2019

Overview

Background

- One of the most common use cases of NGS technologies is to perform experiments comparing two groups of samples (typically disease versus control) to identify **a list of significant (altered) genes**

- This list alone often falls short of providing mechanistic insights into the underlying biology of the disease being studied

- To **reduce the complexity of analysis** while **simultaneously providing great explanatory power**, one can investigate groups of genes that function in the same pathways/gene sets: pathway analysis

Background

Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Motivation

- Utilizing protein-protein interaction information **enhances** pathway enrichment results
  - Successful applications include GNEA, EnrichNet, NetPEA

Liu M, Liberzon A, Kong SW, et al. Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet. 2007;3(6):e96.

Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012;28(18):i451-i457.

Liu L, Wei J, Ruan J. Pathway Enrichment Analysis with Networks. Genes (Basel). 2017;8(10)

- With pathfindR, our aim was likewise to **exploit interaction information** to extract the most relevant pathways

pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks

Ege Ulgen[1]*,   Ozan Ozisik[2] and Osman U. Sezerman[1]

[1]Acıbadem University, Turkey
[2]Yıldız Technical University, Turkey
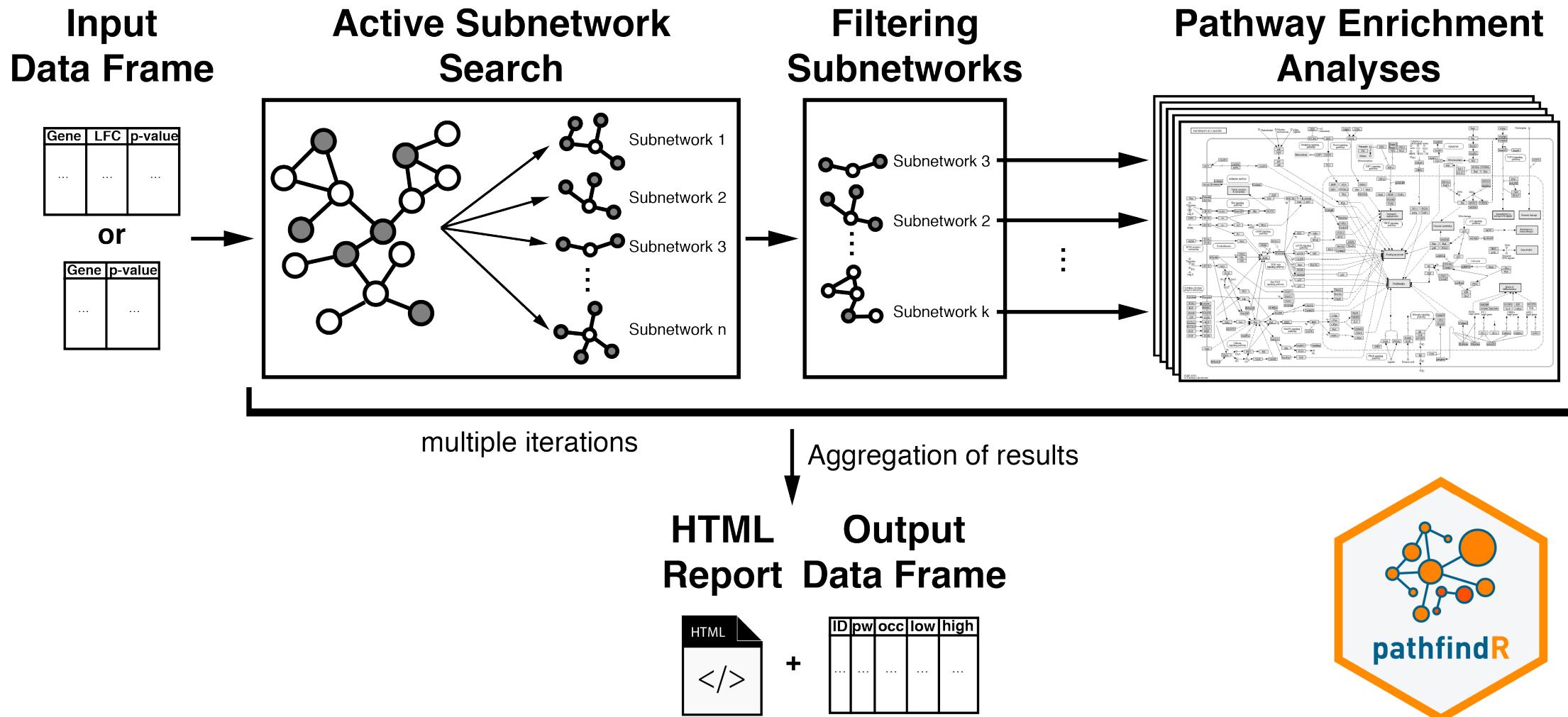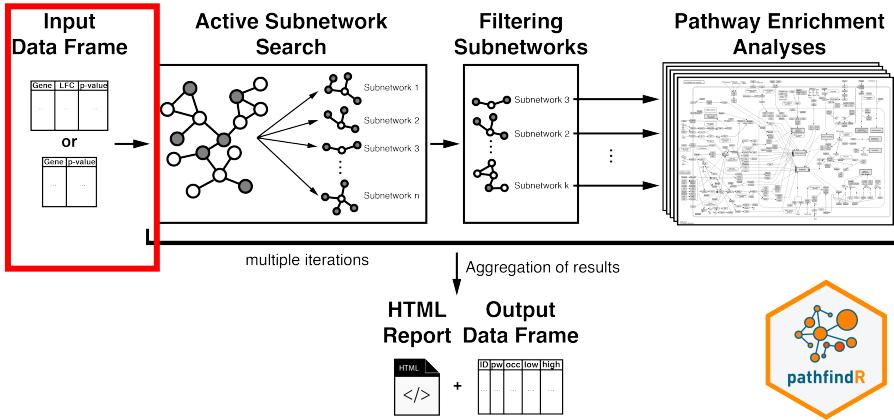
- Using input genes, pathfindR identifies sets of genes that form **active subnetworks** within a protein-protein interaction network

- It then performs **pathway enrichment analyses** on the identified active subnetworks

- Additionally, pathfindR provides functionality to:
  - **Cluster enriched pathways**
  - **Score overall pathway activity**
  - **Visualization of analyses**

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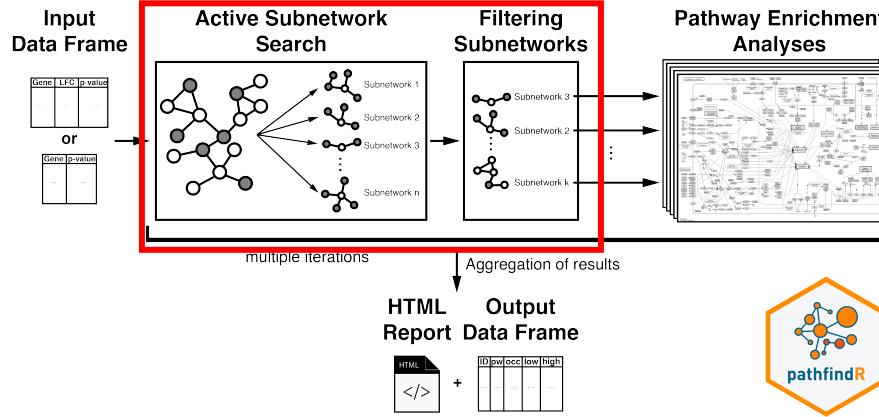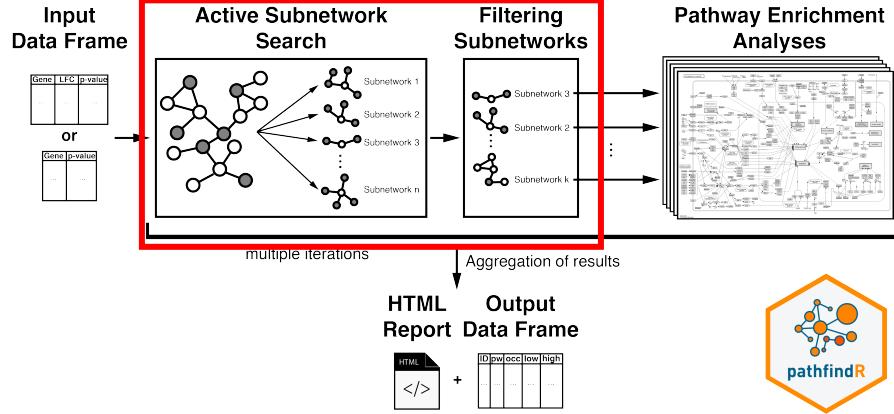
| Gene Symbol	Change Value (OPTIONAL)	Adjusted p-value
FAM110A	-0.6939359	0.0000034
RNASE2	1.3535040	0.0000101
S100A8	1.5448338	0.0000347
S100A9	1.0280904	0.0002263
TEX261	-0.3235994	0.0002263
ARHGAP17	-0.6919330	0.0002708

- Active Subnetwork Search Algorithms:
  - Greedy Algorithm
  - Simulated Annealing
  - Genetic Algorithm

- Available Protein Interaction Networks (PINs):
  - Biogrid
  - GeneMania
  - IntAct
  - KEGG PIN*
  - **Custom PIN**

Active Subnetwork Search

Scoring of Subnetworks

In pathfindR, we followed the scoring scheme that was proposed by Ideker et al., 2002). The p value of each gene is converted to a z score using equation (1), and the score of a subnetwork is calculated using equation (2). In equation (1) Φ^{-1} is the inverse normal cumulative distribution function. In equation (2), A is the set of genes in the subnetwork and k is its cardinality.

$$z_i = \Phi^{-1}(1 - p_i) \quad (1)$$

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \in A} z_i \quad (2)$$
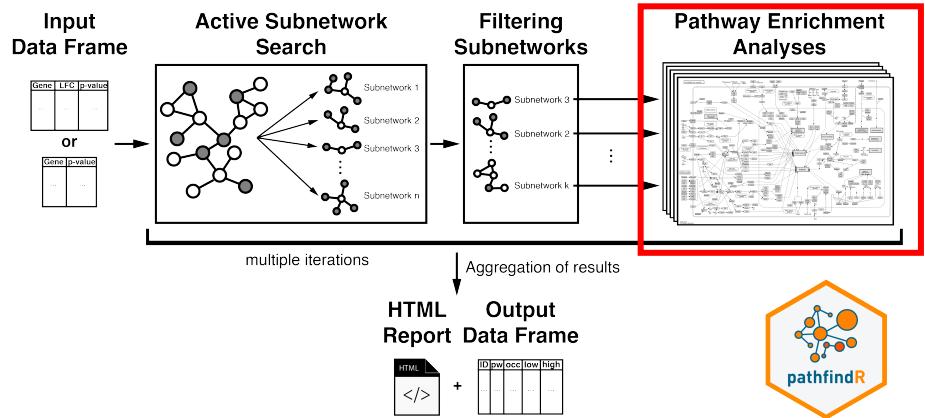
In the same scoring scheme, a Monte Carlo approach is used for the calibration of the scores of subnetworks against a background distribution. Using randomly selected genes, 2,000 subnetworks of each possible size are constructed, and for each possible size, the mean and standard deviation of the score is calculated. These values are used to calibrate the subnetwork score using equation (3).

$$s_A = \frac{(z_A - \mu_k)}{\sigma_k} \quad (3)$$

Subnetwork filtering

An active subnetwork passes the filter if it:
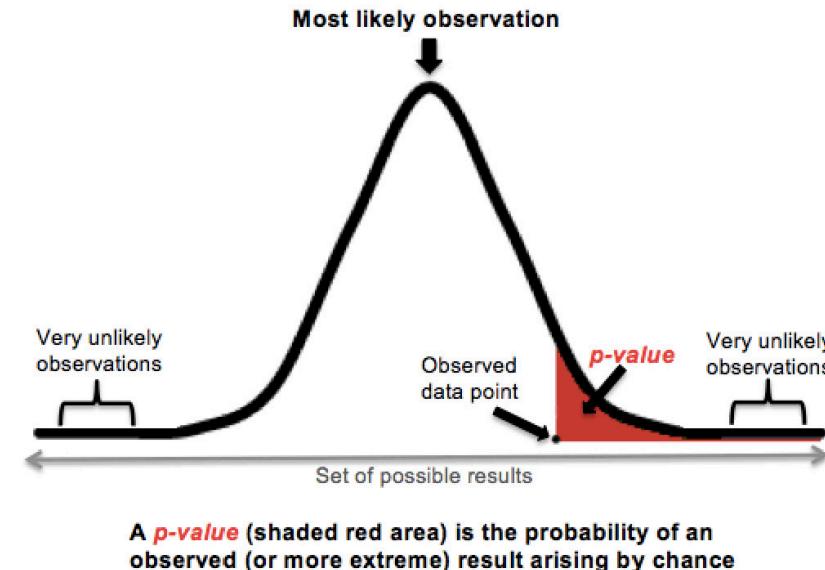
1. has a score larger than the given quantile threshold (default is 0.80) **and**

2. contains at least a specified number of 517 input genes (default is 10).
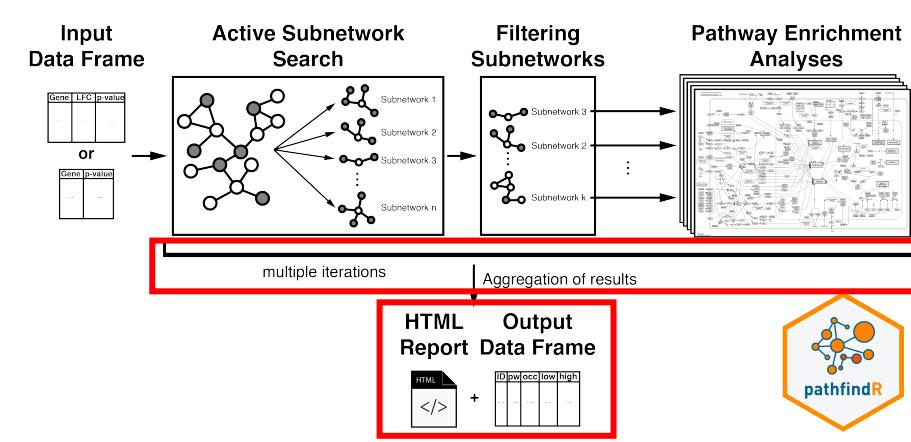
One-sided Hypergeometric Testing

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

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

Most likely observation

Very unlikely observations

Observed data point

p-value

Very unlikely observations

Set of possible results

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

⋮

pathfindR - Results

25 November, 2018

pathfindR-Enrichment results are presented below:

All pathways found to be enriched

A table that lists all pathways found to be enriched as well as lists of up- or down-regulated genes for each pathway. If it was requested, the pathway descriptions (names) are linked to the visualizations of these pathways with differentially-expressed genes colored according to change values.
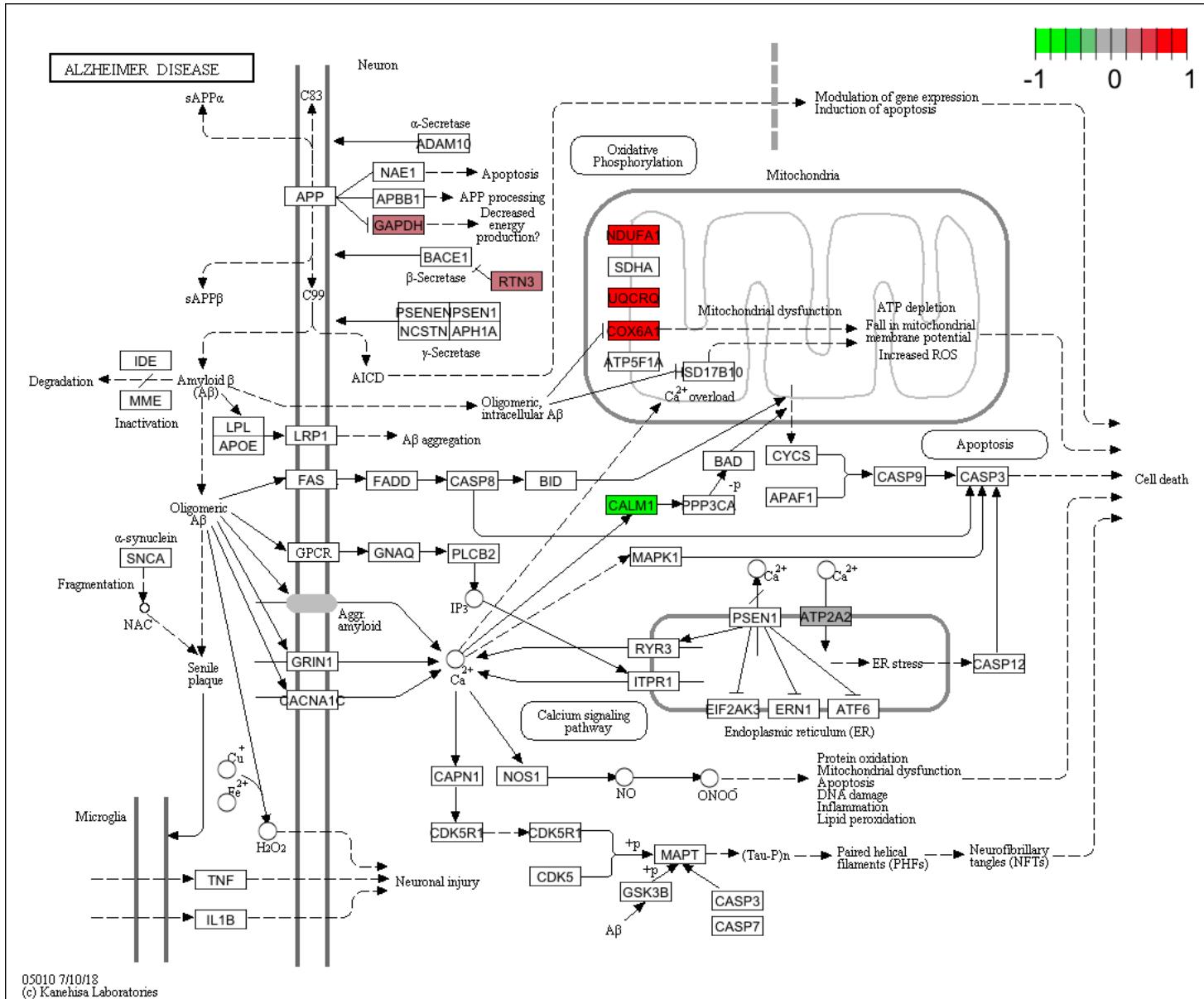
Tables of genes with converted gene symbols and genes without interactions

- A table listing the genes whose symbols (Old Symbol) were converted to aliases (Converted Symbol) that were in the protein-protein interaction network.
- A table listing the input genes for which no interactions in the PIN were found (after the aliases were also checked).

pathfindR - All Enriched Pathways - KEGG
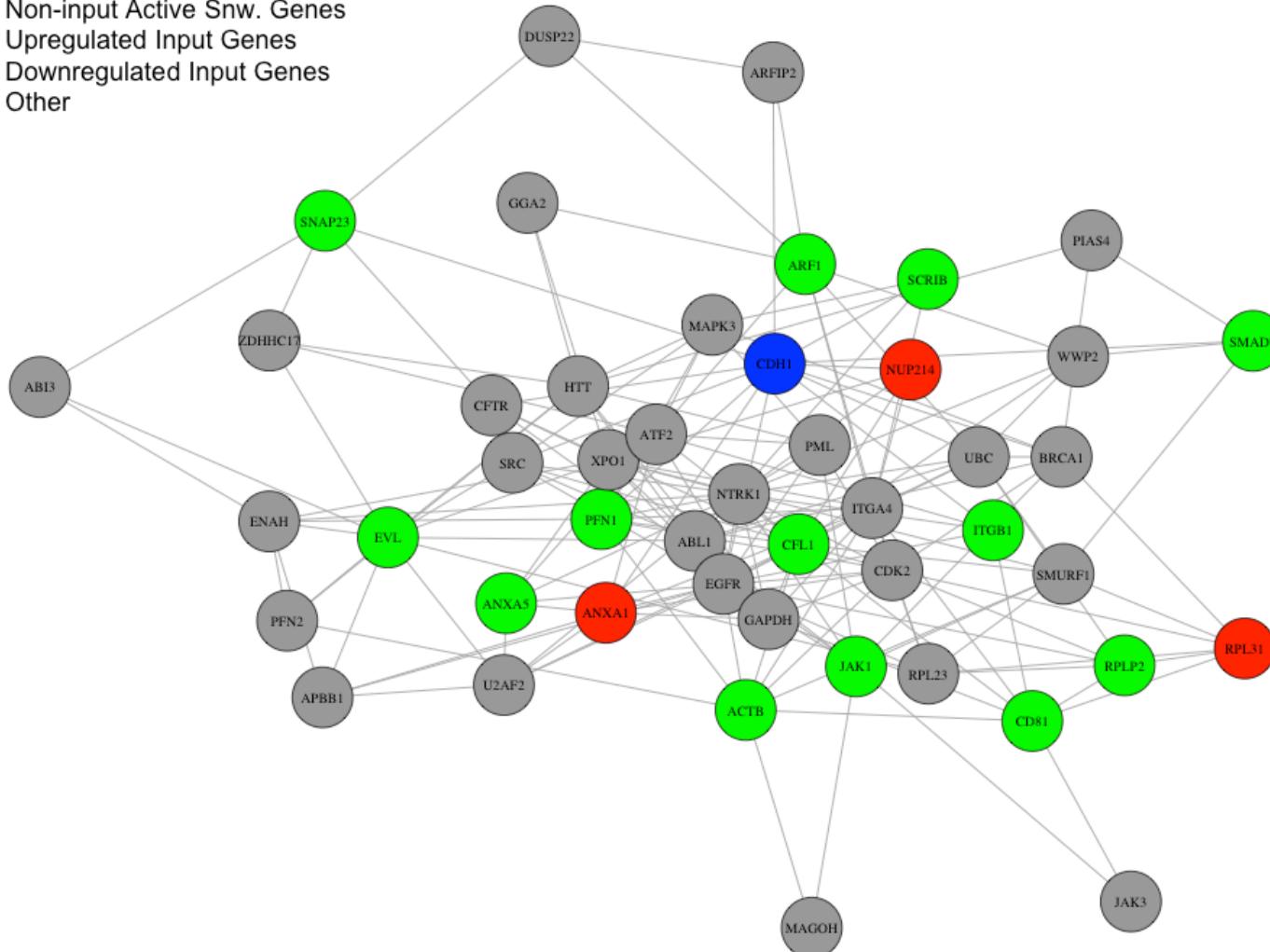
KEGG pathway visualizations were performed via the R Biocondutor package __pathview__ [1].
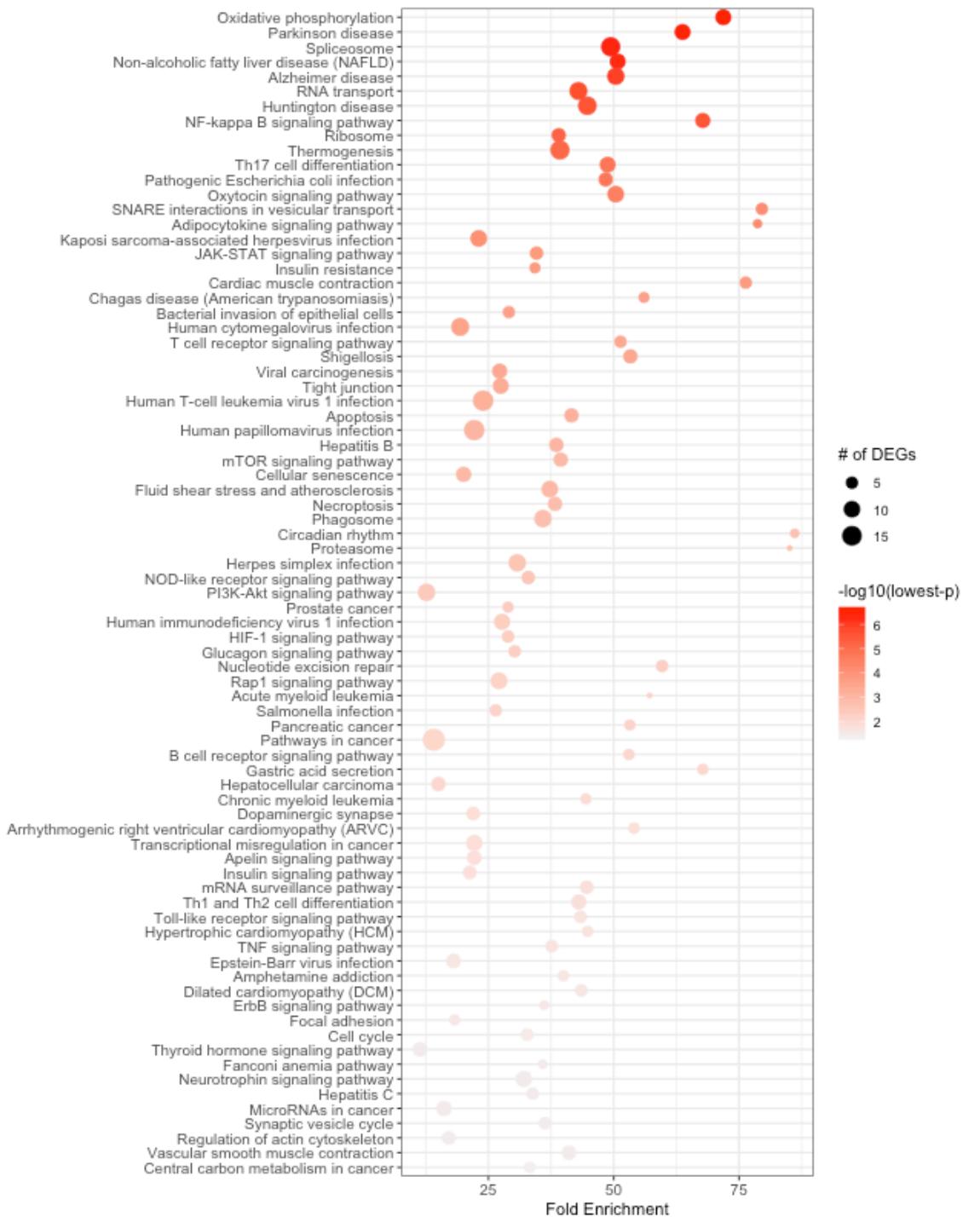
ID	Pathway	Fold_Enrichment	occurrence	lowest_p	highest_p	Up_regulated	Down_regulated
hsa00190	Oxidative phosphorylation	2.52036	10	2.6e-07	2.6e-07	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C, ATP6V1D, ATP6V0E1	ATP6V0E2
hsa05012	Parkinson disease	2.23503	10	3.9e-07	3.9e-07	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	UBE2G1, VDAC1, SLC25A5
hsa03040	Spliceosome	2.90183	10	4.8e-07	4.8e-07	SF3B6, LSM3, BUD31	SNRPB, SF3B2, U2AF2, PUF60, SNU13, DDX23, EIF4A3, HNRNPA1, PCBP1, SRSF8, SRSF5
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	1.78131	10	5.2e-07	5.2e-07	DDIT3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	IKBKB, FASLG
hsa03010	Ribosome	1.47413	10	1.1e-06	6.4e-06	MRPS18C, RPS24, MRPL33, RPL26, RPL31, RPL39	RPLP2
hsa05010	Alzheimer disease	2.16090	10	1.2e-06	1.2e-06	GAPDH, RTN3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	CALM3, CALM1, ATP2A2
hsa03013	RNA transport	2.16360	10	2.1e-06	7.7e-05	NUP214	NUP62, NUP93, RANGAP1, UBE2I, SUMO3, GEMIN4, EIF2S3, EIF2B1, EIF4A3, RNPS1, SRRM1
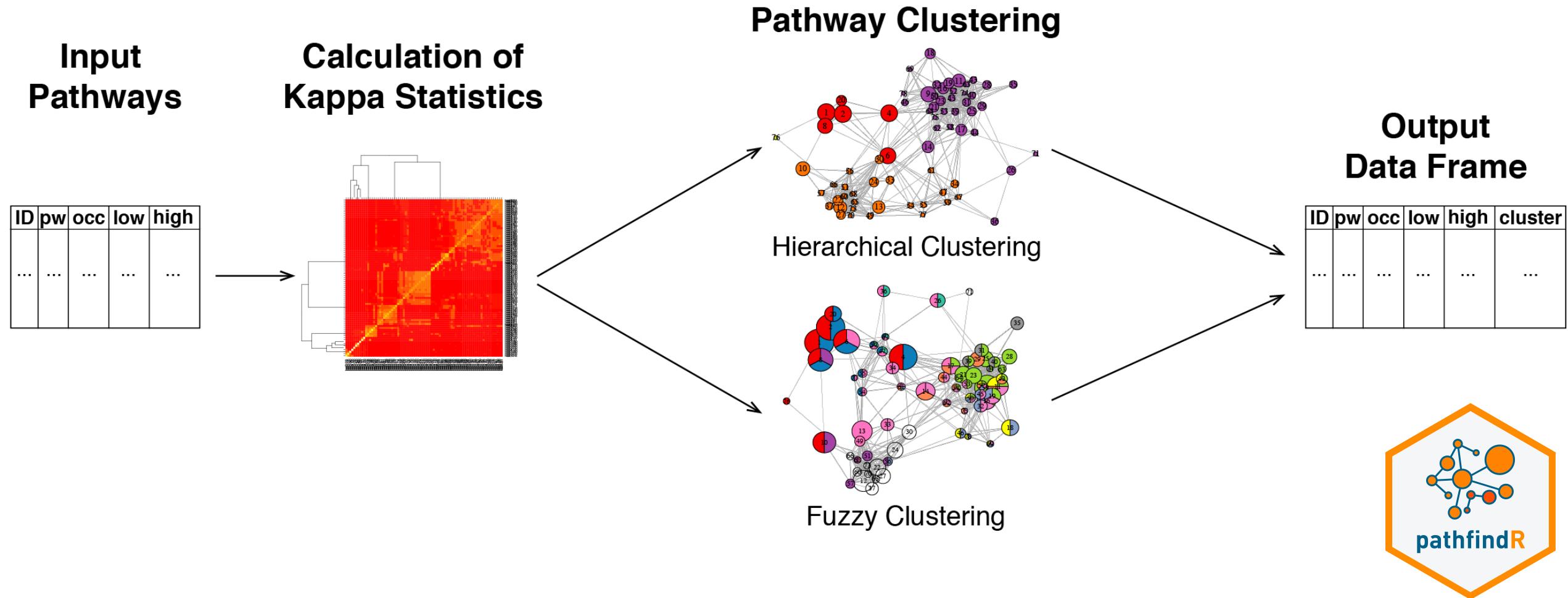
anchoring_junction(3)
Involved Gene Interactions in Biogrid

- Non-input Active Snw. Genes
- Upregulated Input Genes
- Downregulated Input Genes
- Other
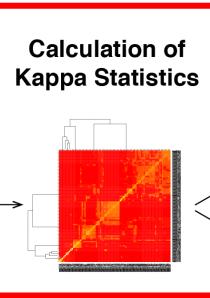
Pathway Clustering Workflow

Using ***1 – kappa similarity*** as distance metric for clustering

Input Pathways

Calculation of Kappa Statistics

Pathway Clustering

Hierarchical Clustering

Fuzzy Clustering

Output Data Frame

ID	pw	occ	low	high	cluster
...	...	...	...	...	...

pathfind**R**

(a)

	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

(b)
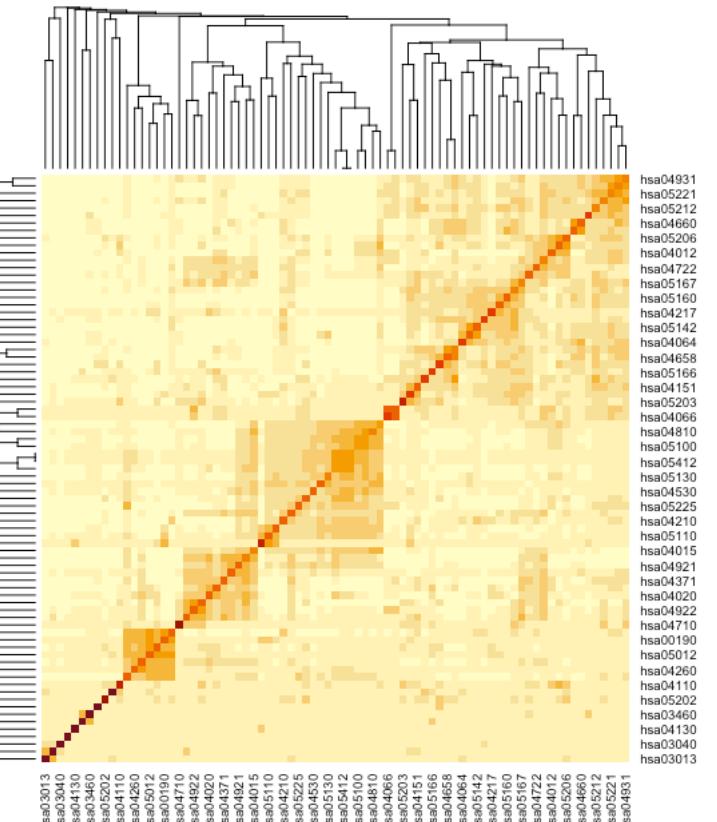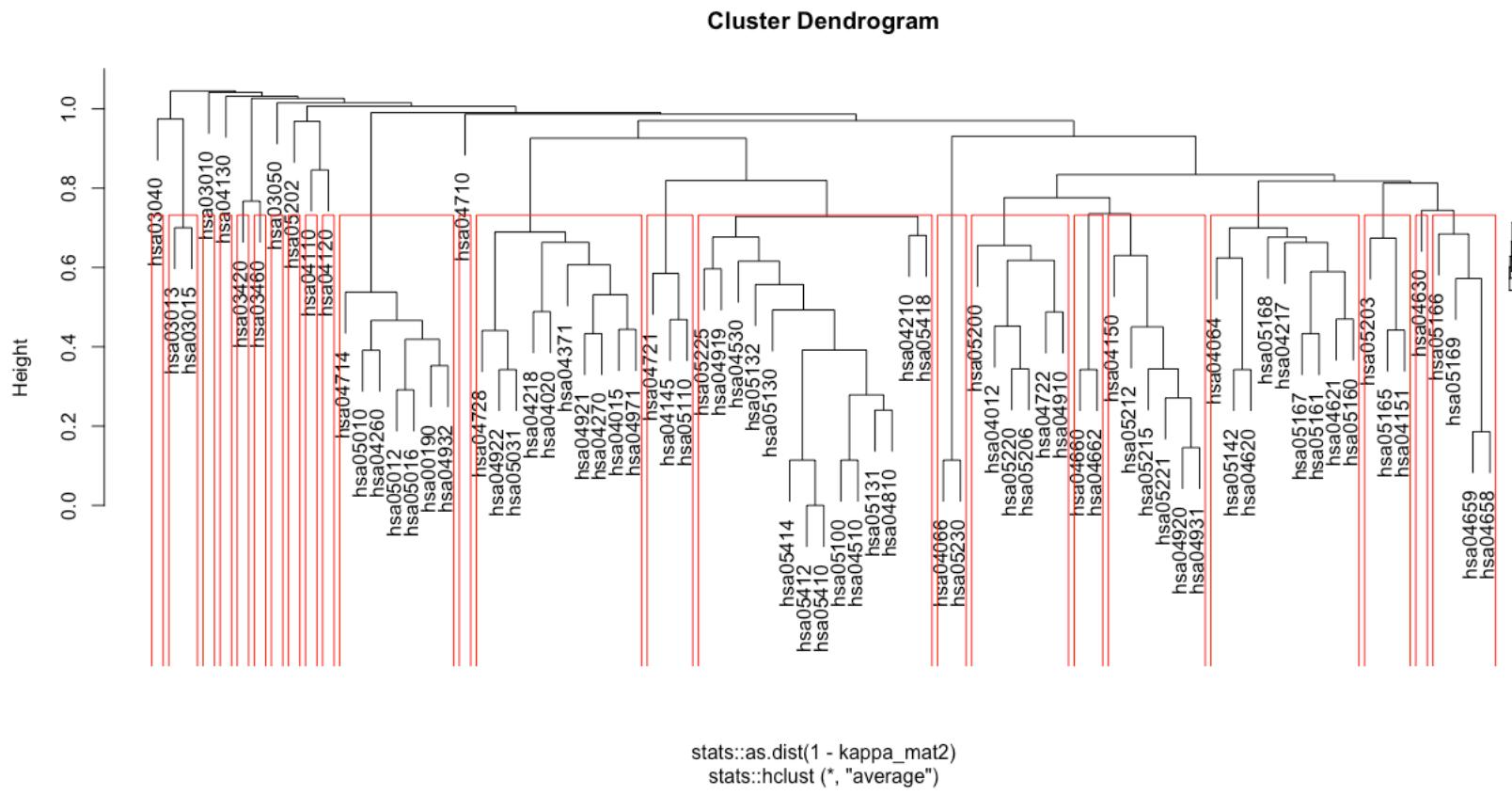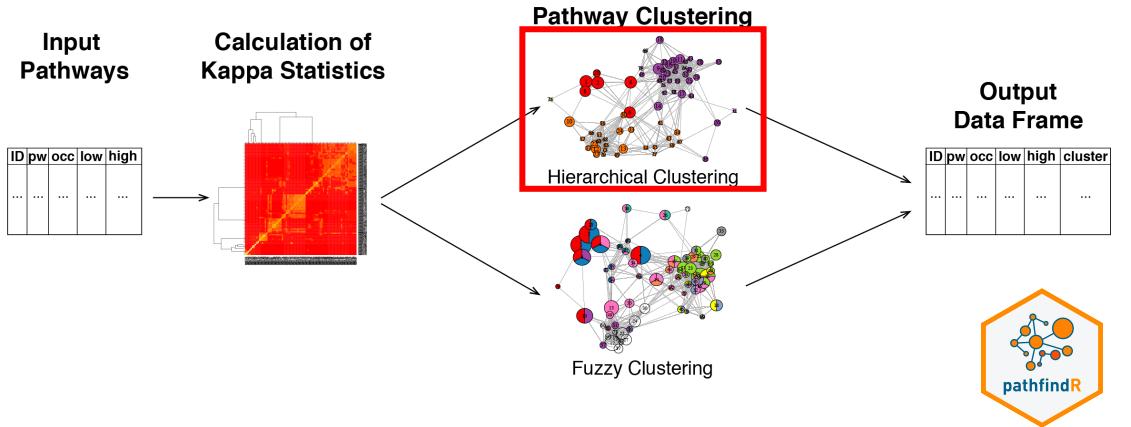
		Gene a		
		1	0	Row total
Gene b	1	$3\ (C_{1,1})$	$1\ (C_{0,1})$	$4\ (C_{1,\cdot})$
	0	$0\ (C_{0,1})$	$2\ (C_{0,0})$	$2\ (C_{0,\cdot})$
Column total		$3\ (C_{\cdot,1})$	$3\ (C_{\cdot,0})$	$6\ (T_{ab})$

$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$
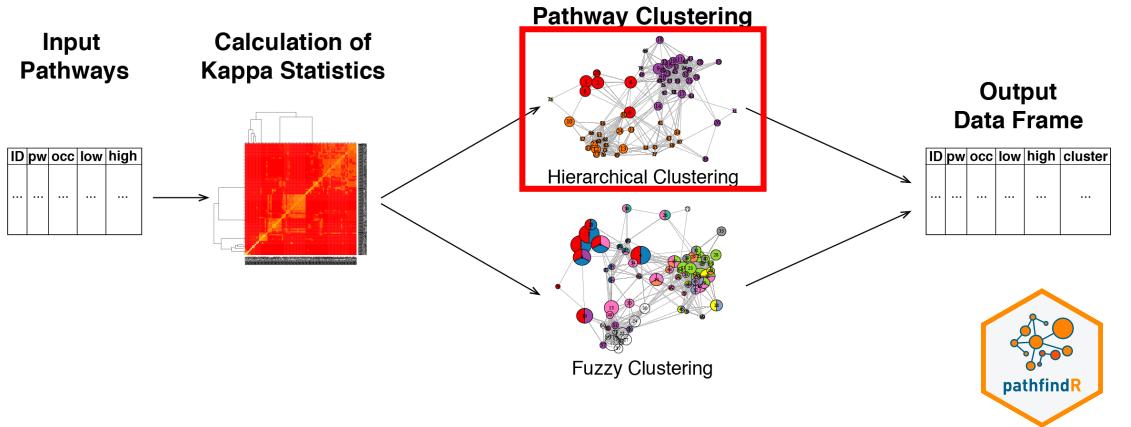
$$A_{ab} = \frac{C_{\cdot,1} \bullet C_{1,\cdot} + C_{\cdot,0} \bullet C_{0,\cdot}}{T_{ab} \bullet T_{ab}} = \frac{3 \bullet 4 + 3 \bullet 2}{6 \bullet 6} = 0.5$$

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.83 - 0.5}{1 - 0.5} = 0.66$$
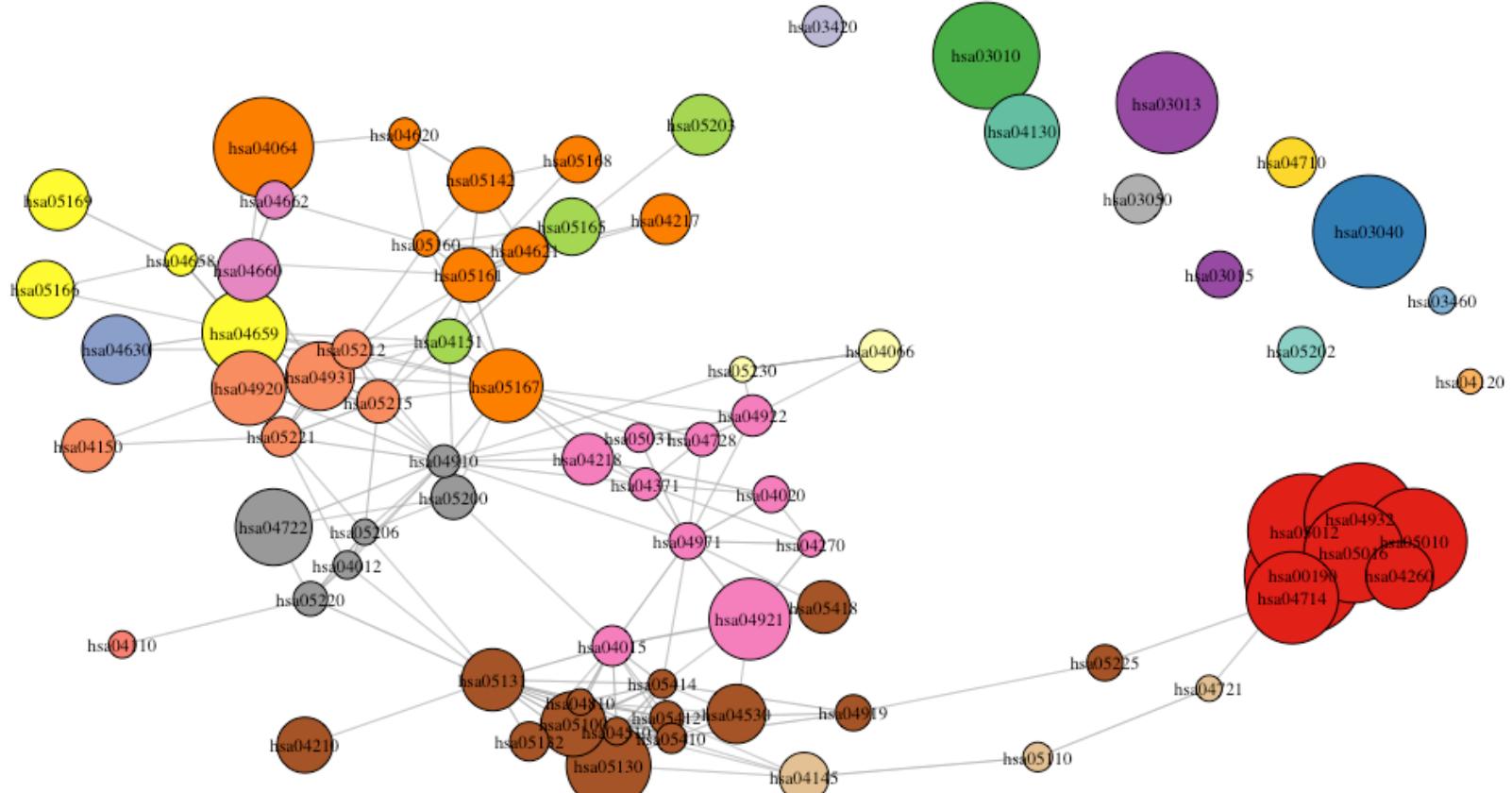
Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):R183.
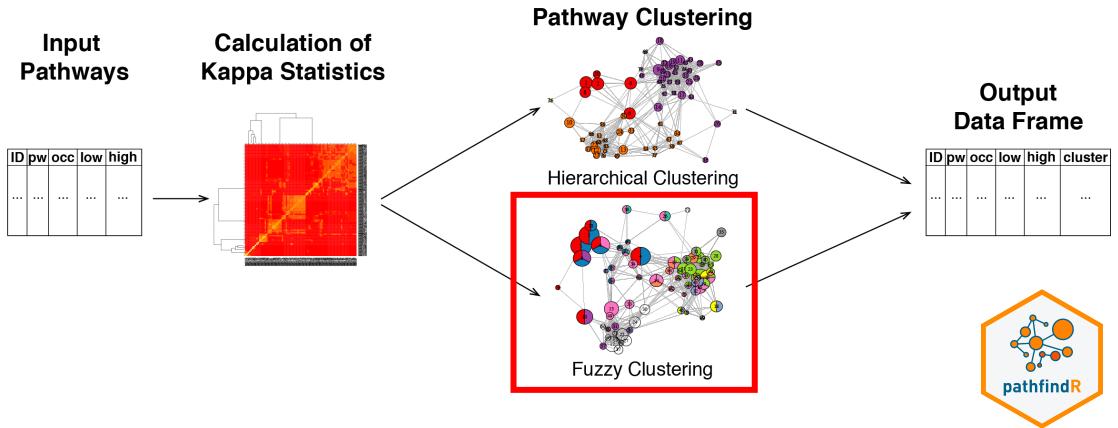
The optimal number of clusters is automatically determined by maximizing the average silhouette width

No links shown for kappa < 0.35 (default)
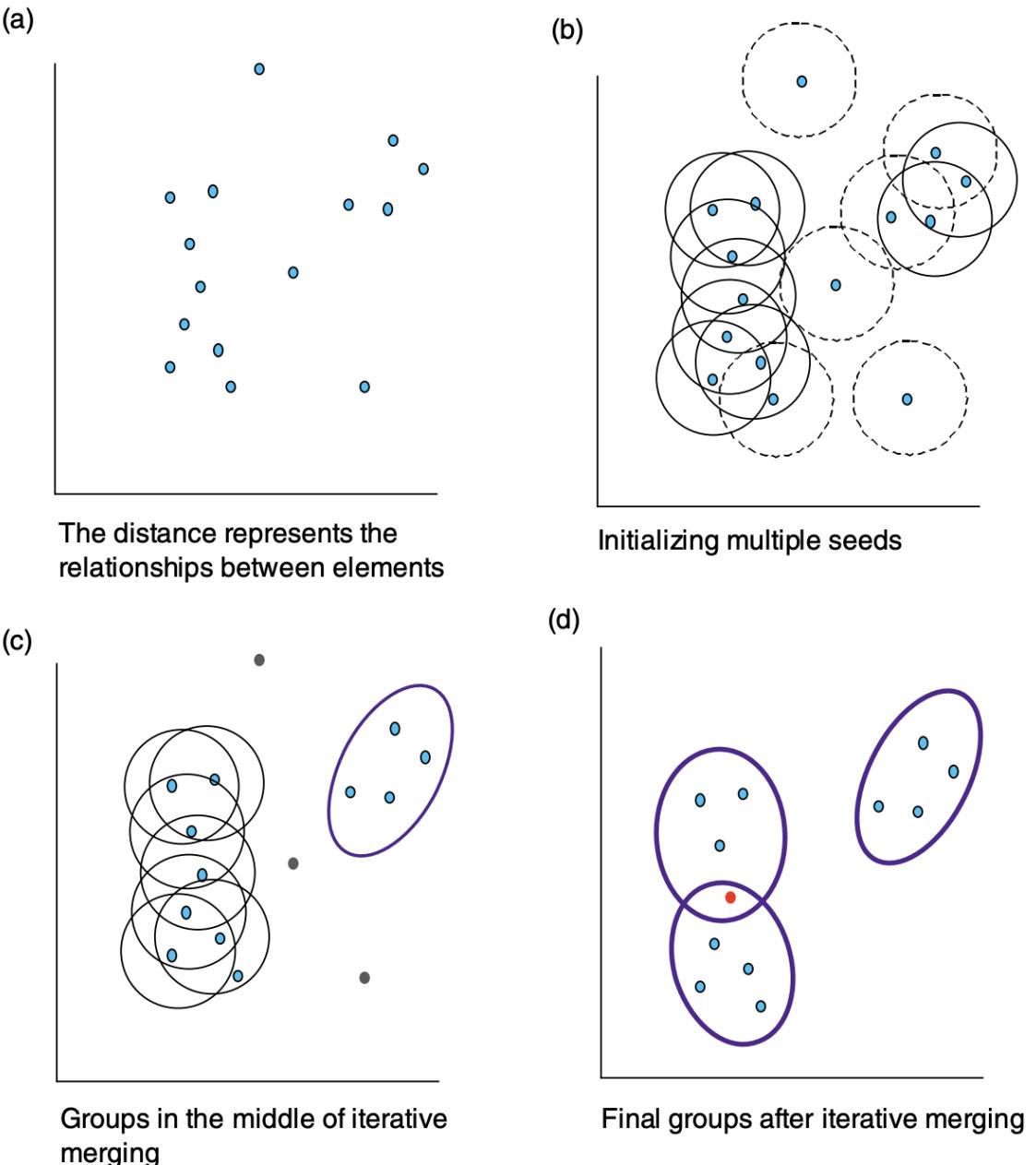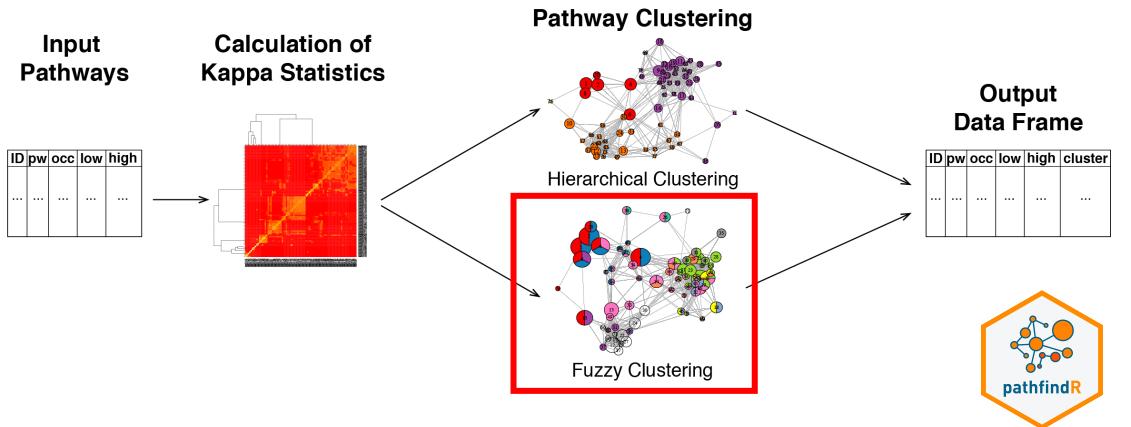
The heuristic fuzzy partition algorithm
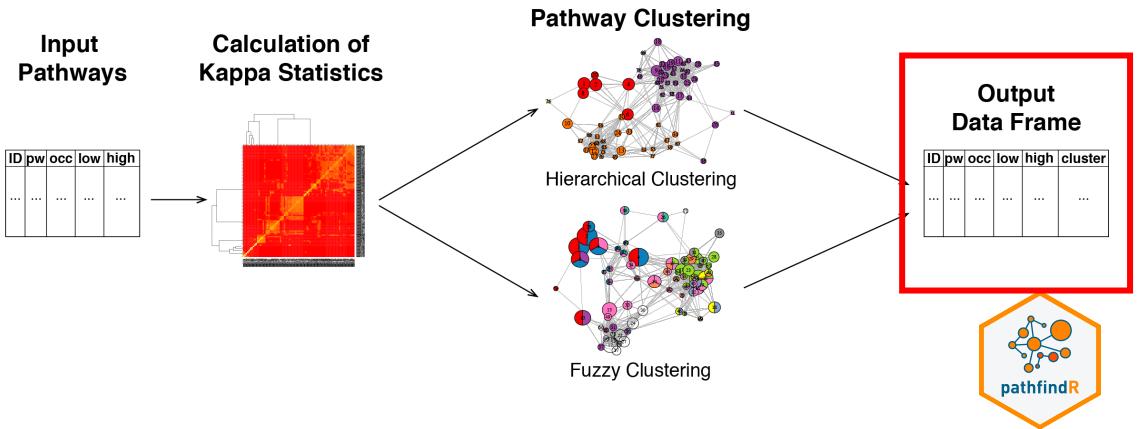
(a)

The distance represents the relationships between elements

(b)

Initializing multiple seeds

(c)

Groups in the middle of iterative merging

(d)

Final groups after iterative merging

Using
1 – kappa similarity
as distance metric for clustering

Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):R183.

Input Pathways

ID	pw	occ	low	high
...	...	...	...	...

Calculation of Kappa Statistics

Pathway Clustering

Hierarchical Clustering

Fuzzy Clustering

Output Data Frame

ID	pw	occ	low	high	cluster
...	...	...	...	...	...

pathfind**R**

No links shown for kappa < 0.35 (default)

Representative pathway selection
For each cluster, the representative gene set is chosen as the one with the lowest p value (default)
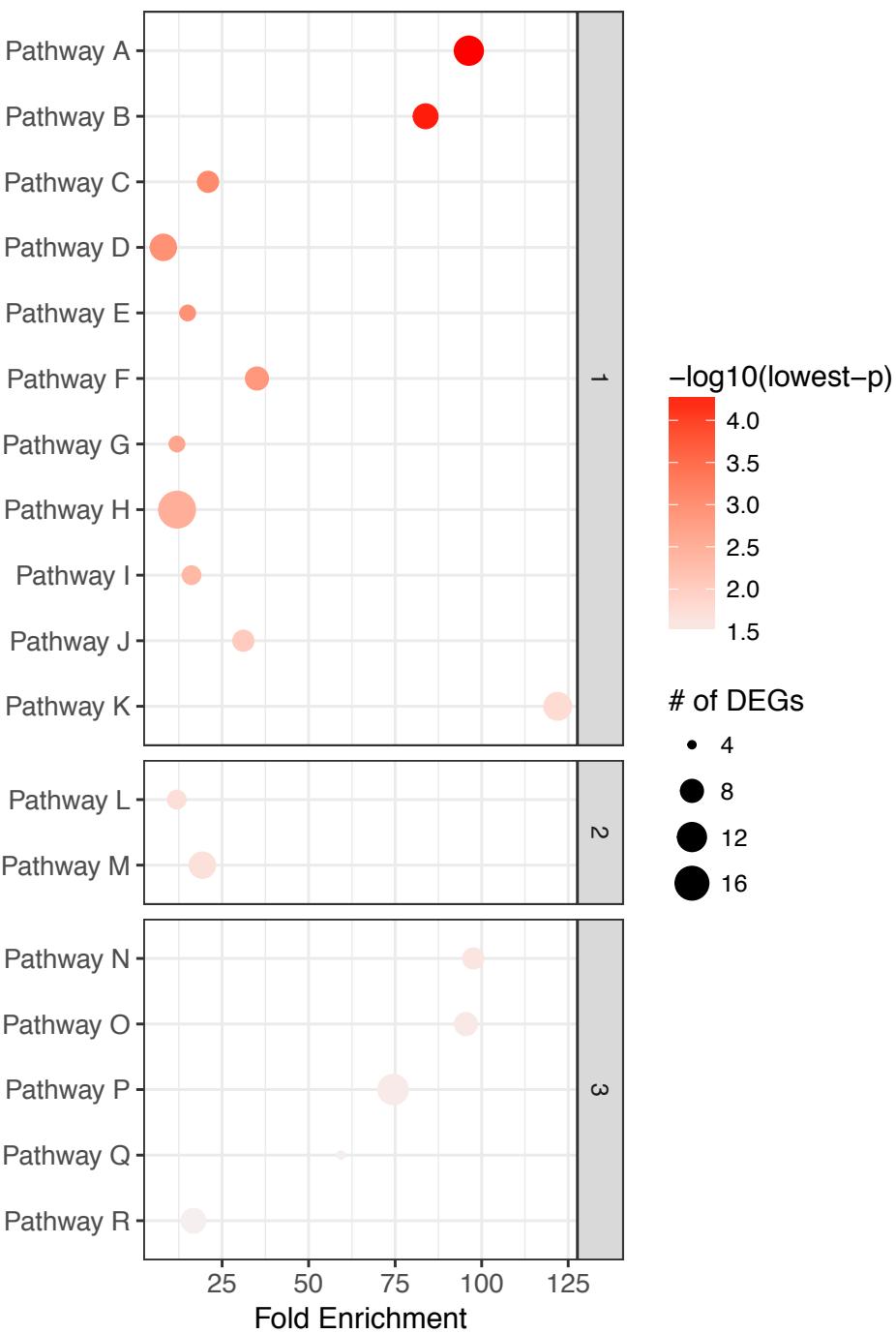
Note that this is an **ad hoc** decision and different approaches may be used:
- Highest fold enrichment
- The most biologically meaningful, etc.
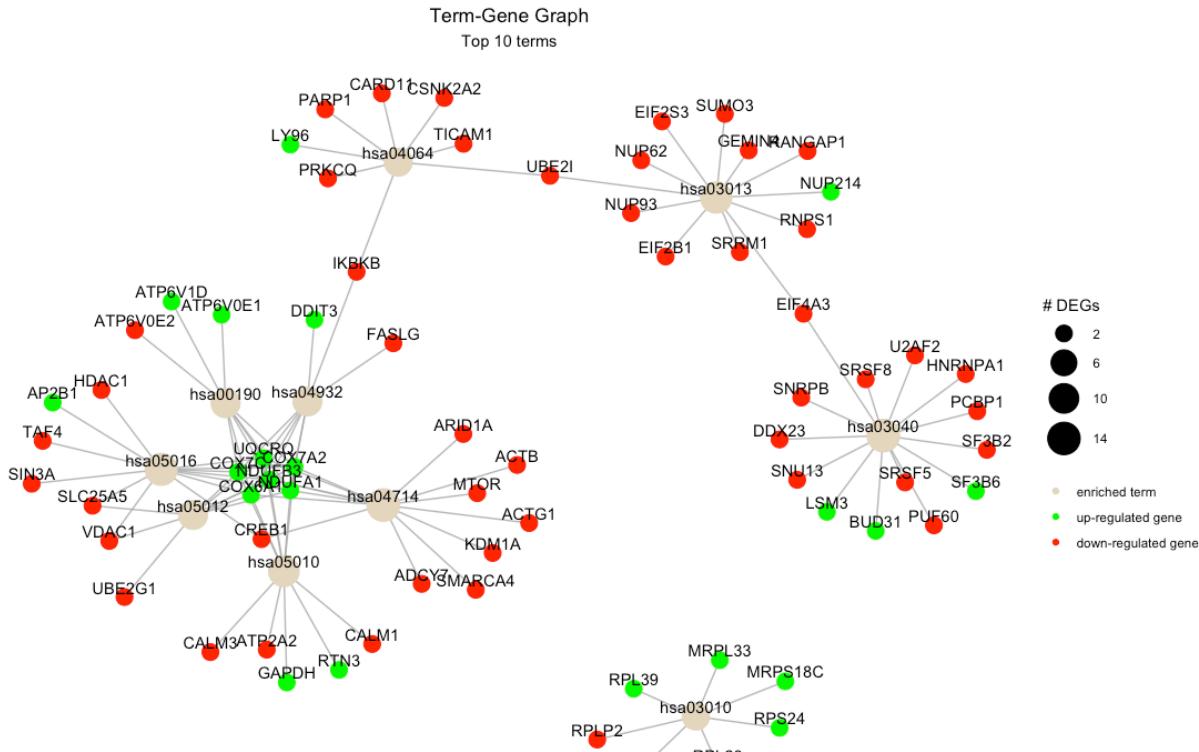
| ID	Pathway	Fold_Enrichment	occurrence	lowest_p	highest_p	Up_regulated	Down_regulated	Cluster	Status
hsa00190	Oxidative phosphorylation	71.863	10	2.61E-07	2.61E-07	NDUFB3, NDUFA1, COX7C	ATP6V0E2	1	Representative
hsa05012	Parkinson's disease	63.727	10	3.88E-07	3.88E-07	UQCRQ, COX6A1, COX7A2	VDAC1, UBE2G1	1	Member
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	50.79	10	5.19E-07	5.19E-07	DDIT3,COX6A1, COX7A2	FASLG, IKBKB	2	Representative

⋮

Term-Gene Graph
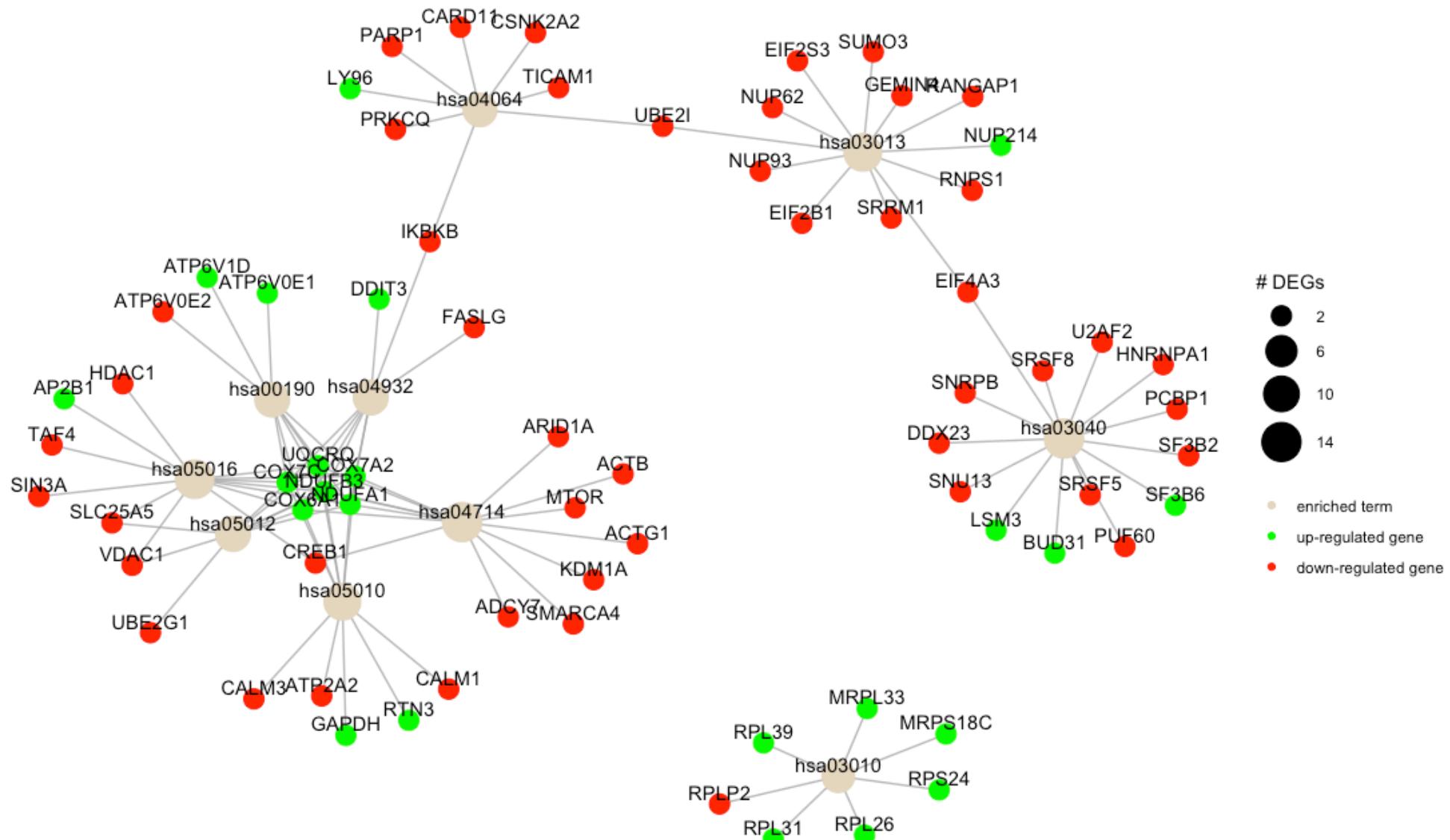
- Graph representation of enriched terms and related genes
  - Do different terms share common genes?
  - Is there a distinct set of genes that are related to a given term?

- Nodes:
  - Enriched terms (beige)
  - Up-regulated genes (green) or
  - Down-regulated genes (red)
- Edges:
  - Term-gene: the given term (pathway or gene set) involves the gene
- Sizes of term nodes are proportional to either:
  - the number of genes (default)
  - the $-\log_{10}$(p value)

Term-Gene Graph
Top 10 terms

Pathway Scoring

For a set of pathways $P = \{P_1, P_2, ..., P_n\}$, where each P_i contains a set of genes, i.e. $P_i = \{g_1, g_2, ..., g_k\}$, the pathway score matrix *PS* is defined as:
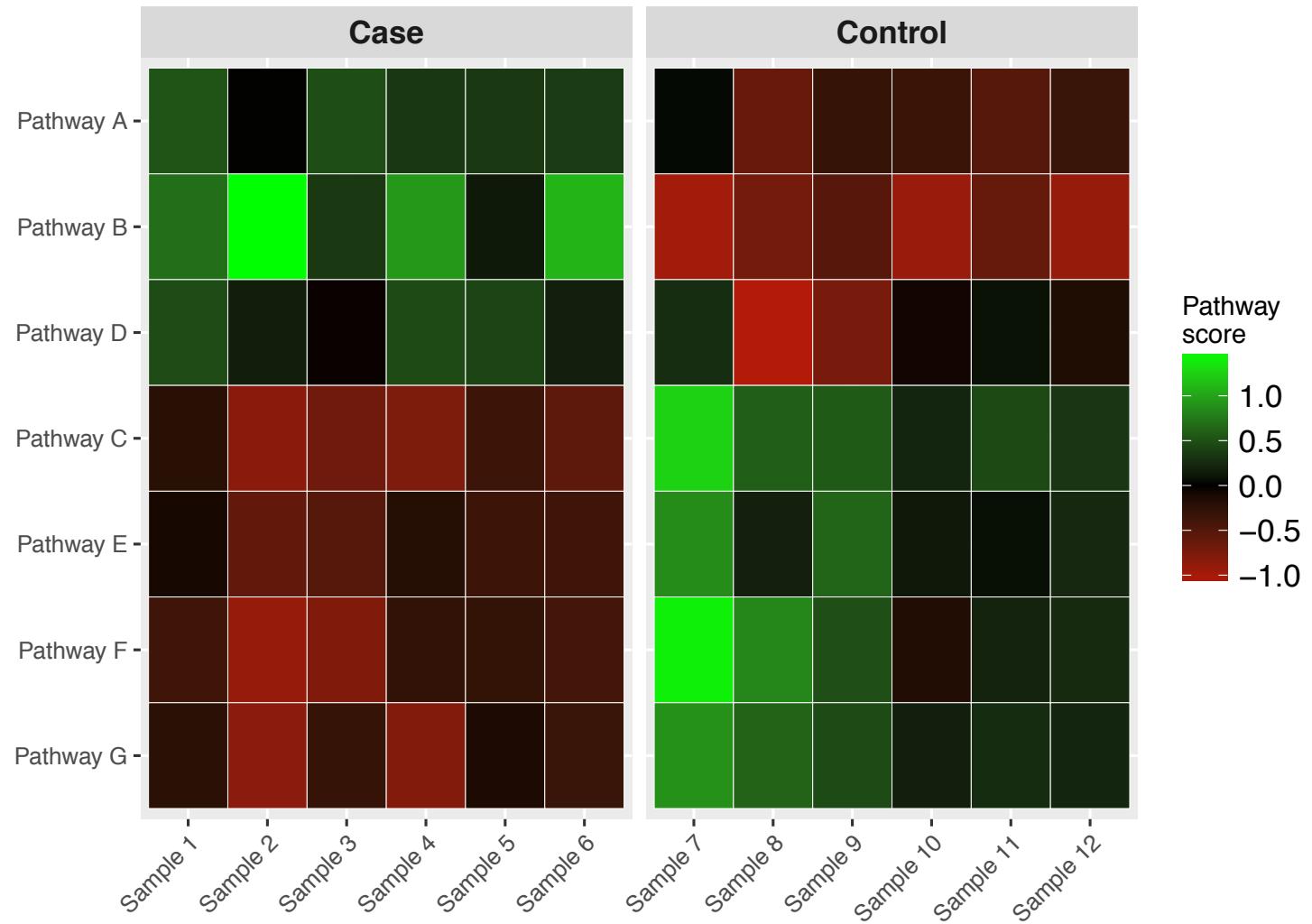
$PS_{p,s} = \frac{1}{k} \sum_{g \in P_p} GS_{g,s}$ for each pathway *p* and for each sample *s*.

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

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

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 deviaton of the expression values for the gene.

Pathway Scoring

Demonstration

Important Note

- For the active subnetwork search component to work(i.e., in order to run pathfindR), the user must:

1. have *Java* installed and

2. *path/to/java* must be in the *PATH* environment variable

   - For Windows users, to configure the *PATH* environment variable see: https://github.com/egeulgen/pathfindR/wiki/Installation#configuration-of-java-on-windows

Demo – I – Installation (CRAN release version)

Installation – Bioconductor Dependencies

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

Installation – pathfindR

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

OR from DockerHub

```
# pull image for latest release
docker pull egeulgen/pathfindr:latest
# pull image for specific version (e.g. 1.3.0)
docker pull egeulgen/pathfindr:1.3.0
```

pathfindR

Demo – II – Installation (Dev. version)

From GitHub

```
install.packages("devtools") # if you have not installed "devtools" package
devtools::install_github("egeulgen/pathfindR")
```

From DockerHub

```
# pull image for development version
docker pull egeulgen/pathfindr:dev
```

pathfindR

pathfindR: Pathway Enrichment Analysis Utilizing Active Subnetworks

Pathway enrichment analysis enables researchers to uncover mechanisms underlying the phenotype. pathfindR is a tool for pathway enrichment analysis utilizing active subnetworks. It identifies active subnetworks in a protein-protein interaction network using user-provided a list of genes. It performs pathway enrichment analyses on the identified subnetworks. pathfindR also offers functionalities to cluster enriched pathways and identify representative pathways and to score the pathways per sample. The method is described in detail in Ulgen E, Ozisik O, Sezerman OU. 2018. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. bioRxiv. <doi:10.1101/272450>.

Version:	1.3.0
Depends:	R (≥ 3.4), knitr, pathview
Imports:	AnnotationDbi, DBI, doParallel, foreach, rmarkdown, org.Hs.eg.db, ggplot2, fpc, grDevices, igraph
Published:	2018-11-20
Author:	Ege Ulgen, Ozan Ozisik
Maintainer:	Ege Ulgen <egeulgen at gmail.com>
BugReports:	https://github.com/egeulgen/pathfindR/issues
License:	MIT + file LICENSE
URL:	https://github.com/egeulgen/pathfindR
NeedsCompilation:	no
SystemRequirements:	Java JVM 1.8
Citation:	pathfindR citation info
Materials:	NEWS
CRAN checks:	pathfindR results

Downloads:

Reference manual:	pathfindR.pdf
Vignettes:	pathfindR - Step-by-Step (Manual) Execution of the pathfindR Workflow
	pathfindR - An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks
Package source:	pathfindR_1.3.0.tar.gz
Windows binaries:	r-devel: pathfindR_1.3.0.zip, r-release: pathfindR_1.3.0.zip, r-oldrel: pathfindR_1.3.0.zip
OS X binaries:	r-release: not available, r-oldrel: not available
Old sources:	pathfindR archive

CRAN release 1.3.0

Demo – III – Pathway Enrichment (w\ wrapper)

```r
library(pathfindR)

RA_demo <- run_pathfindR(RA_input)

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

pathfindR

Demo – IV – Pathway Clustering

```r
# hierarchical clustering (default)
RA_demo_clu <- cluster_pathways(RA_demo)


# fuzzy clustering
RA_demo_clu <- cluster_pathways(RA_demo,
                                method = "fuzzy")
```

pathfindR

Demo – V – Term-Gene Graph

```r
term_gene_graph(RA_demo) # top 10 terms (default)


# Graph using representative pathways
# selecting "Representative" pathways for clear visualization
RA_representative <- RA_clustered[RA_clustered $Status == "Representative", ]
term_gene_graph(RA_representative,
                num_terms = NULL, # to plot using all terms
                use_names = TRUE) # use pw names instead of IDs
```

pathfindR

Demo – VI – Pathway Scoring

```r
# 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)
```

pathfindR

Resources

- Tutorial on Biostars:
  - https://www.biostars.org/p/322415/

- Vignettes
  - https://cran.r-project.org/web/packages/pathfindR/vignettes/

- 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

pathfindR