pathfindR

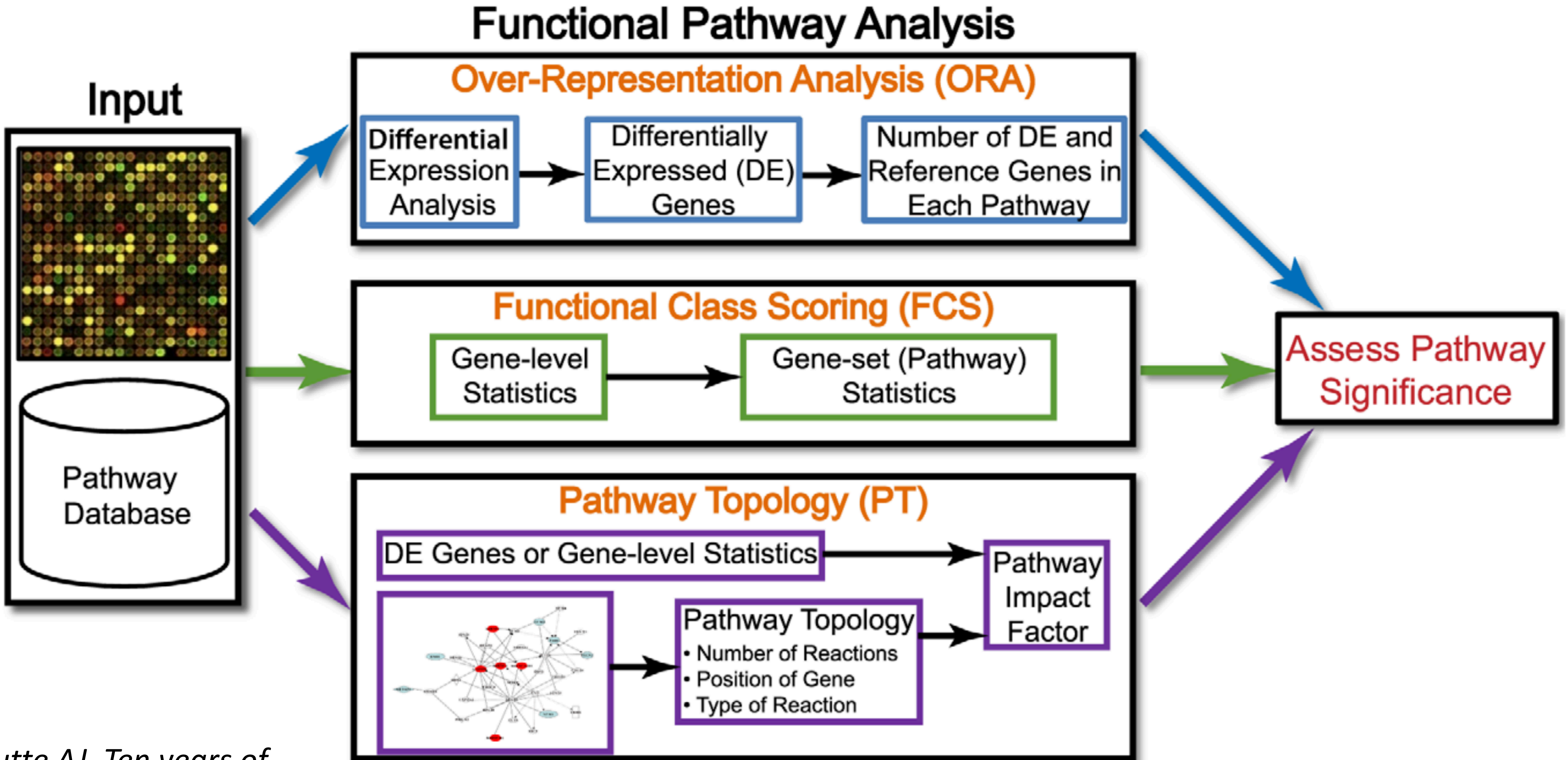
Enrichment Analysis Utilizing Active Subnetworks

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: **enrichment 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** enrichment results
  - Previous successful applications include GNEA, EnrichNet, NetPEA, PANOGA*

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)

Bakir-gungor B, Egemen E, Sezerman OU. PANOGA: a web server for identification of SNP-targeted pathways from genome-wide association study data. Bioinformatics. 2014;30(9):1287-9.

- With pathfindR, our aim was likewise to **exploit interaction information** to extract the most relevant gene sets (of pathways/gene ontology terms/transcription factor target genes, miRNA target genes etc.)

** pathfindR was developed based on PANOGA: a previous approach developed by our group for genome-wide association studies* frontiers in Genetics

METHODS ARTICLE

Front. Genet., 25 September 2019 | https://doi.org/10.3389/fgene.2019.00858

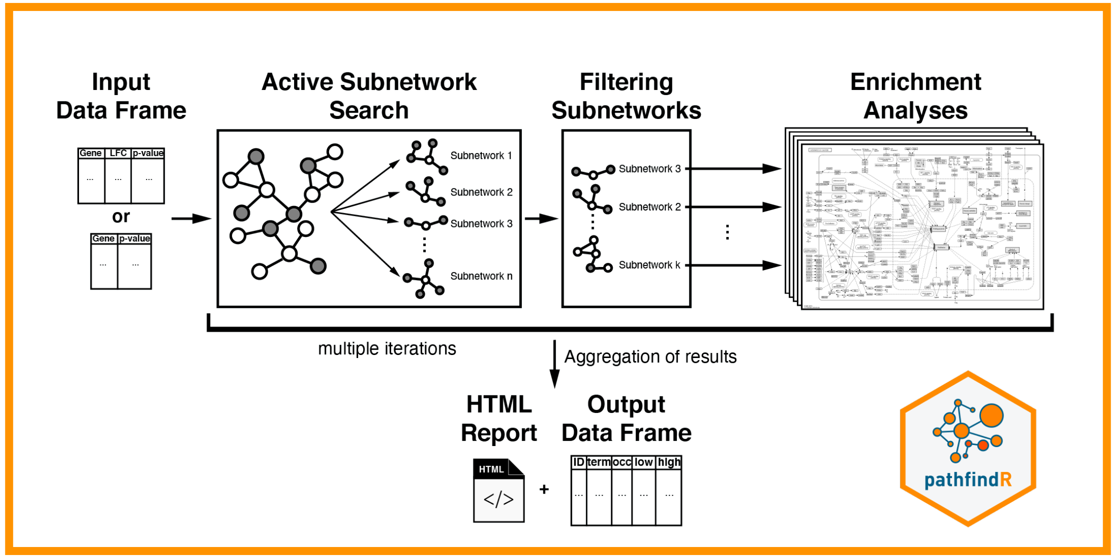
Check for updates

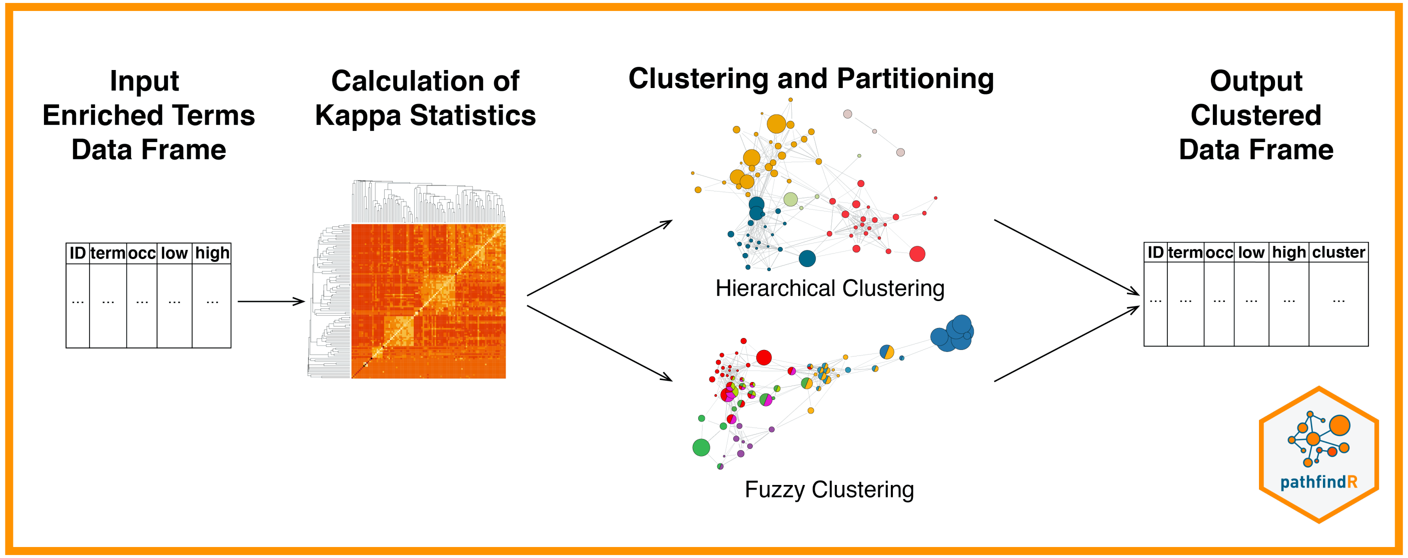
pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks

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

[1]Department of Biostatistics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

[2]Department of Computer Engineering, Electrical & Electronics Faculty, Yildiz Technical University, Istanbul, Turkey





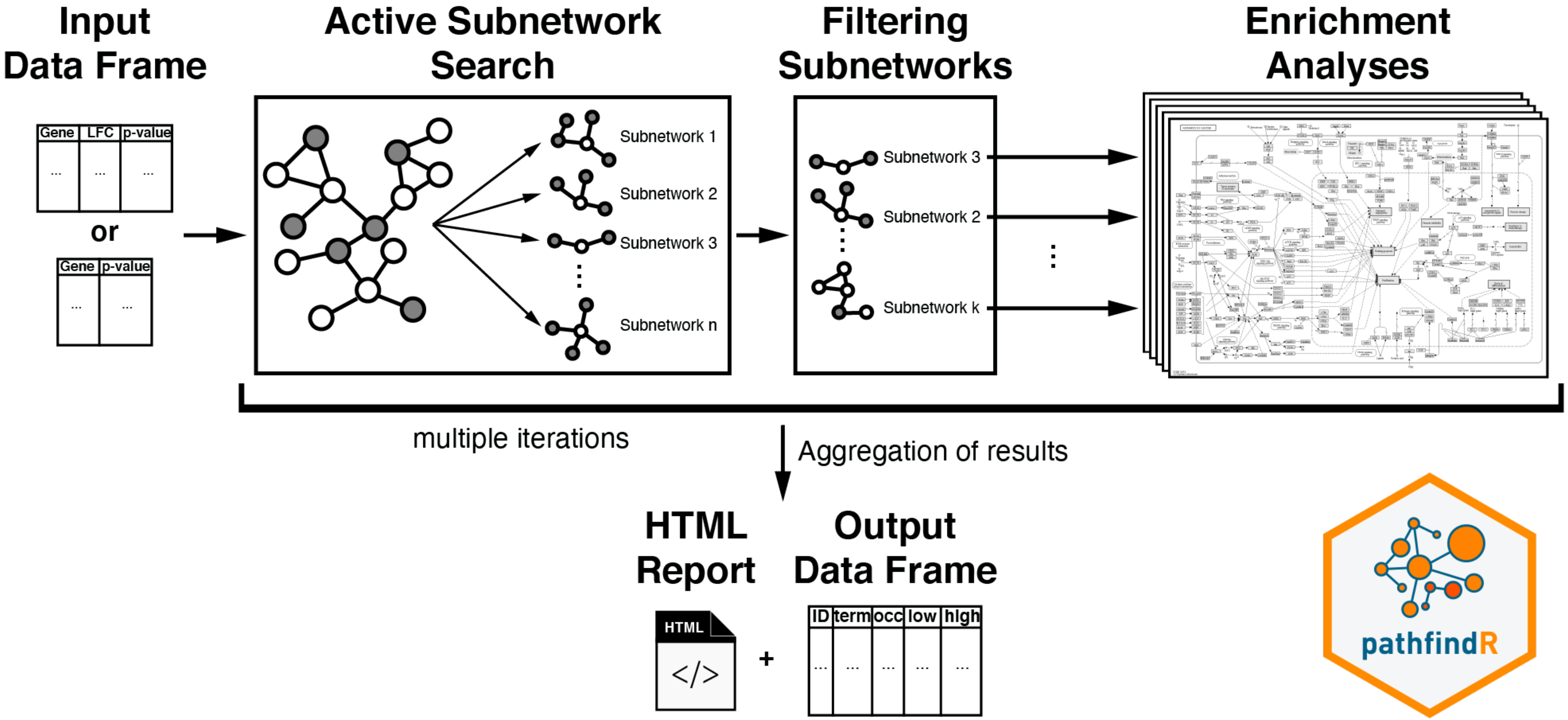
- Using input genes, pathfindR identifies sets of genes that form **active subnetworks** within a protein-protein interaction network

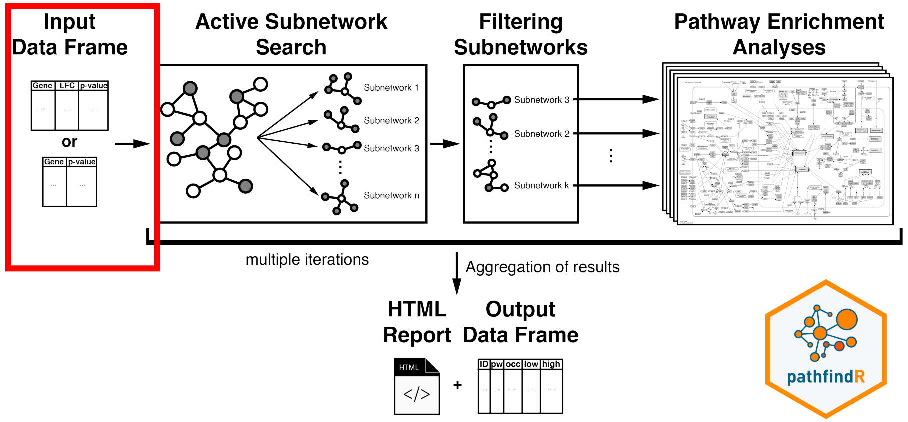
  An active subnetwork can be defined as a group of interconnected genes in a PIN that predominantly consists of significantly altered genes.

- It then performs **enrichment analyses** on the identified active subnetworks (see above diagram)
- Additionally, pathfindR provides functionality to:
  - **Cluster enriched terms** (see above diagram)
  - Calculate **agglomerated score per term activity per subject**
  - Create various **visualizations** of the analysis

pathfindR

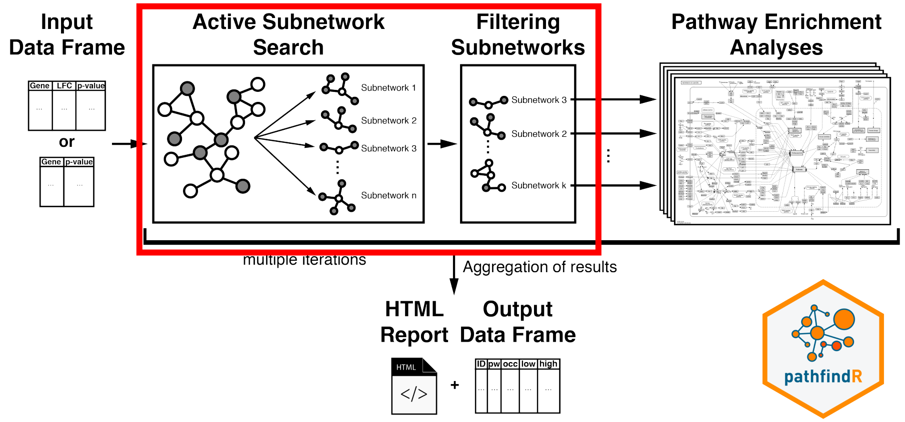
Active Snw.-oriented Enrichment Workflow





Gene Symbol	Change Value (OPTIONAL)	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

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.

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

$$z_A = \frac{1}{\sqrt{k}} \sum_{i\ A} z_i \tag{2}$$

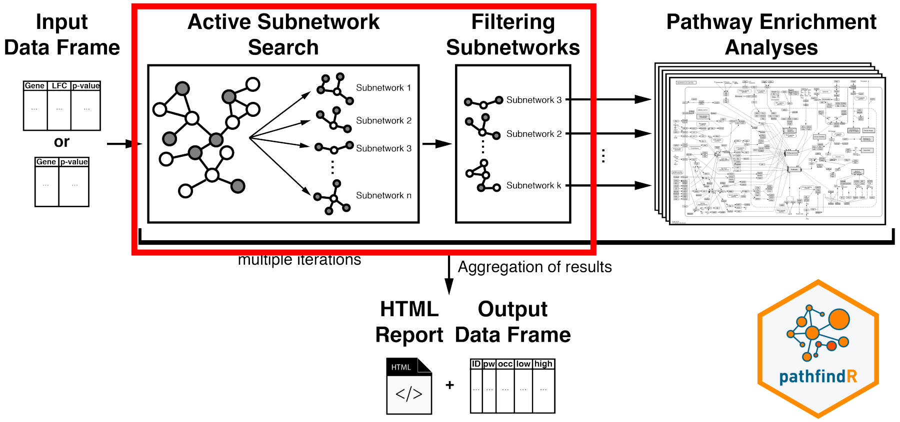
In the same scoring scheme, a Monte Carlo approach is used for the calibration of the scores of subnetworks against a background distribution. Using randomly selected genes, 2,000 subnetworks of each possible size are constructed, and for each possible size, the mean and standard deviation of the score is calculated. These values are used to calibrate the subnetwork score using equation (3).

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

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

- Available Protein Interaction Networks (PINs):
  - Biogrid*
  - STRING
  - GeneMania
  - IntAct
  - KEGG PIN
  - mmu_STRING (M.musculus)
  - **Custom PIN** (path/to/PIN)

**default options for pathfindR*



Active Subnetwork Search

Subnetwork filtering

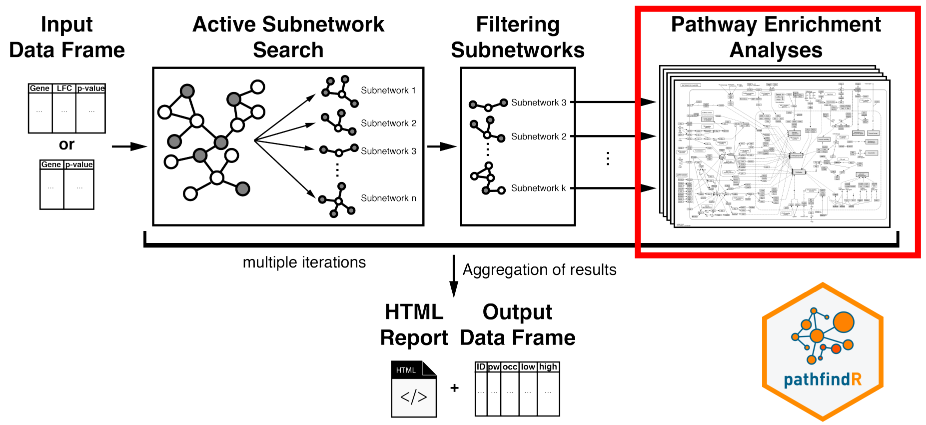
An active subnetwork passes the filter if it:

1. has a score larger than the given quantile threshold (default is 0.80) **and**
2. contains at least a specified proportion of input genes (default is 0.02).

Choice of Active Subnetwork Search Method

- In pathfindR, we use multiple subnetworks obtained via the chosen active subnetwork search algorithm
- We then filter the subnetworks and perform enrichment on the genes of each of these subnetworks separately and the enrichment results are aggregated later
- For this approach, the default greedy algorithm is sufficient and fast

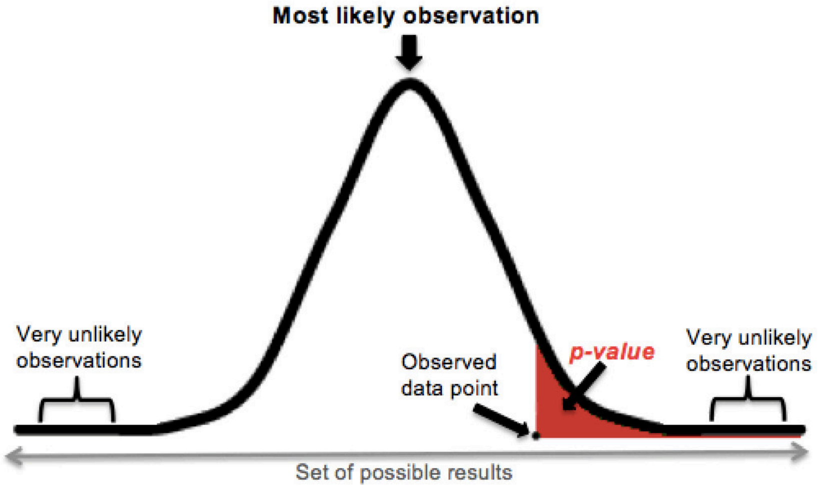
- If the user decides to use the single highest scoring active subnetwork for the enrichment process, they are encouraged to consider greedy algorithm with greater depth, simulated annealing or genetic algorithm



One-sided Hypergeometric Testing

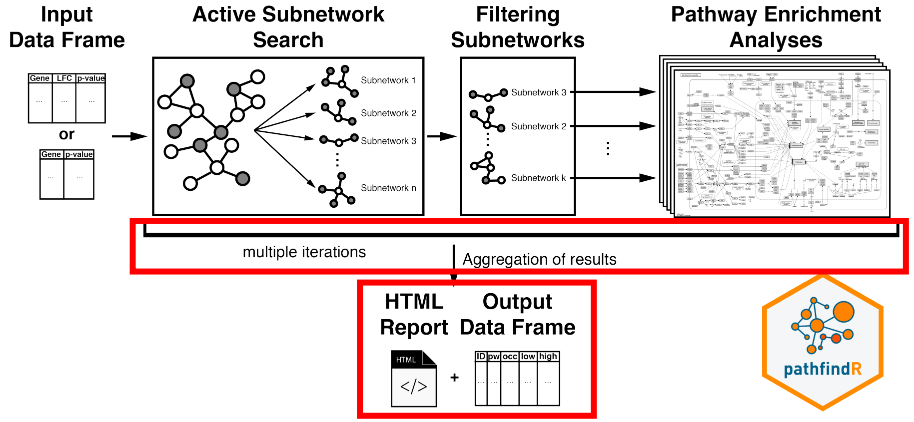
- Available gene sets/pathways:
  - KEGG*
  - Reactome
  - BioCarta
  - Gene Ontology gene sets
    - GO – All (i.e., GO-BP + GO-CC + GO-MF)
    - GO – BP
    - GO – CC
    - GO – MF
  - mmu_KEGG (M.musculus KEGG)
  - **Custom gene sets/pathways**

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



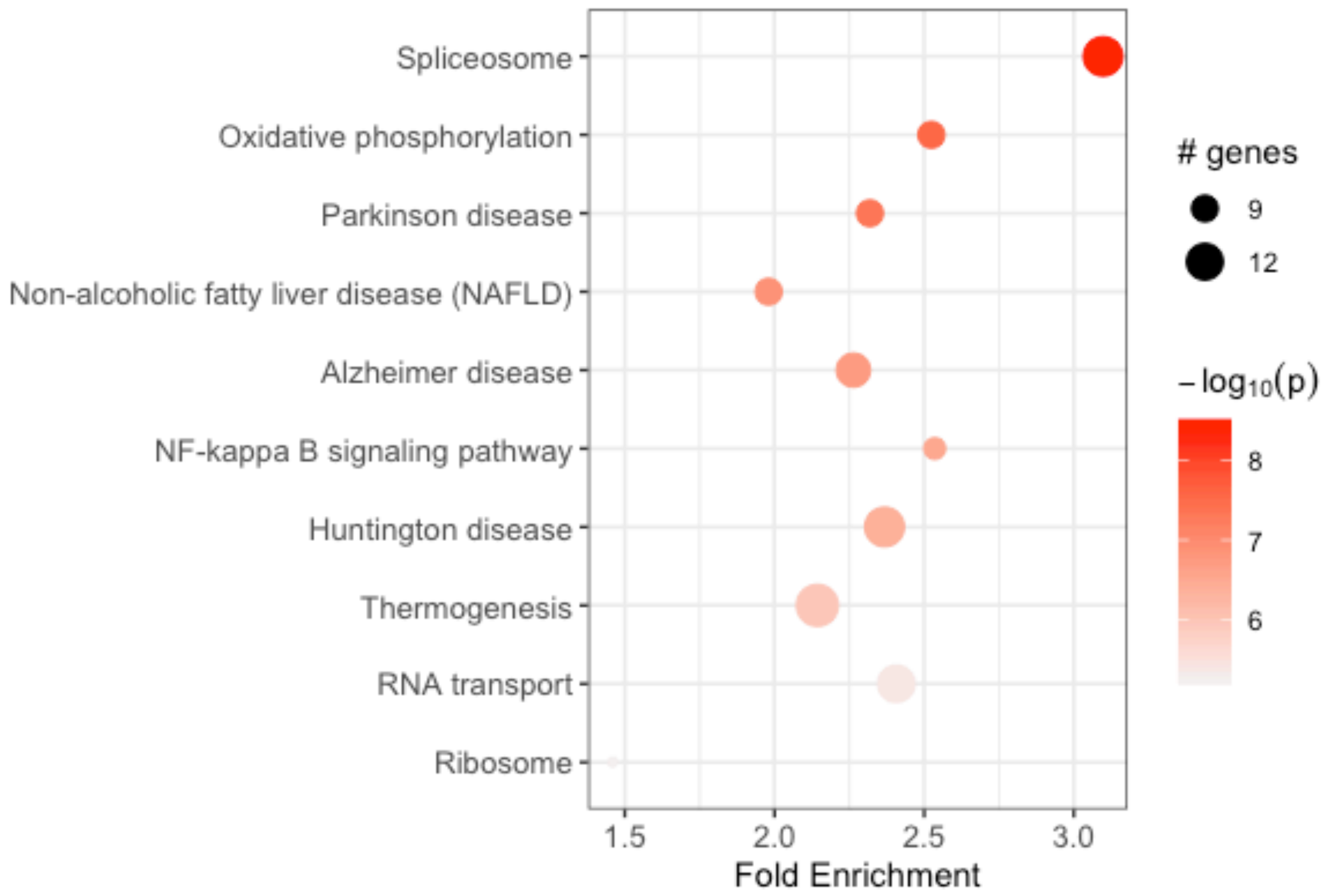
A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

**default gene sets for pathfindR*



ID	Term_Description	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

⋮



Spliceosome
Oxidative phosphorylation
Parkinson disease
Non-alcoholic fatty liver disease (NAFLD)
Alzheimer disease
NF-kappa B signaling pathway
Huntington disease
Thermogenesis
RNA transport
Ribosome
1.5
2.0
2.5
3.0
Fold Enrichment
genes
9
12
$-\log_{10}(p)$
8
7
6

pathfindR - Results

01 November, 2019

pathfindR-Enrichment results are presented below:

All terms found to be enriched

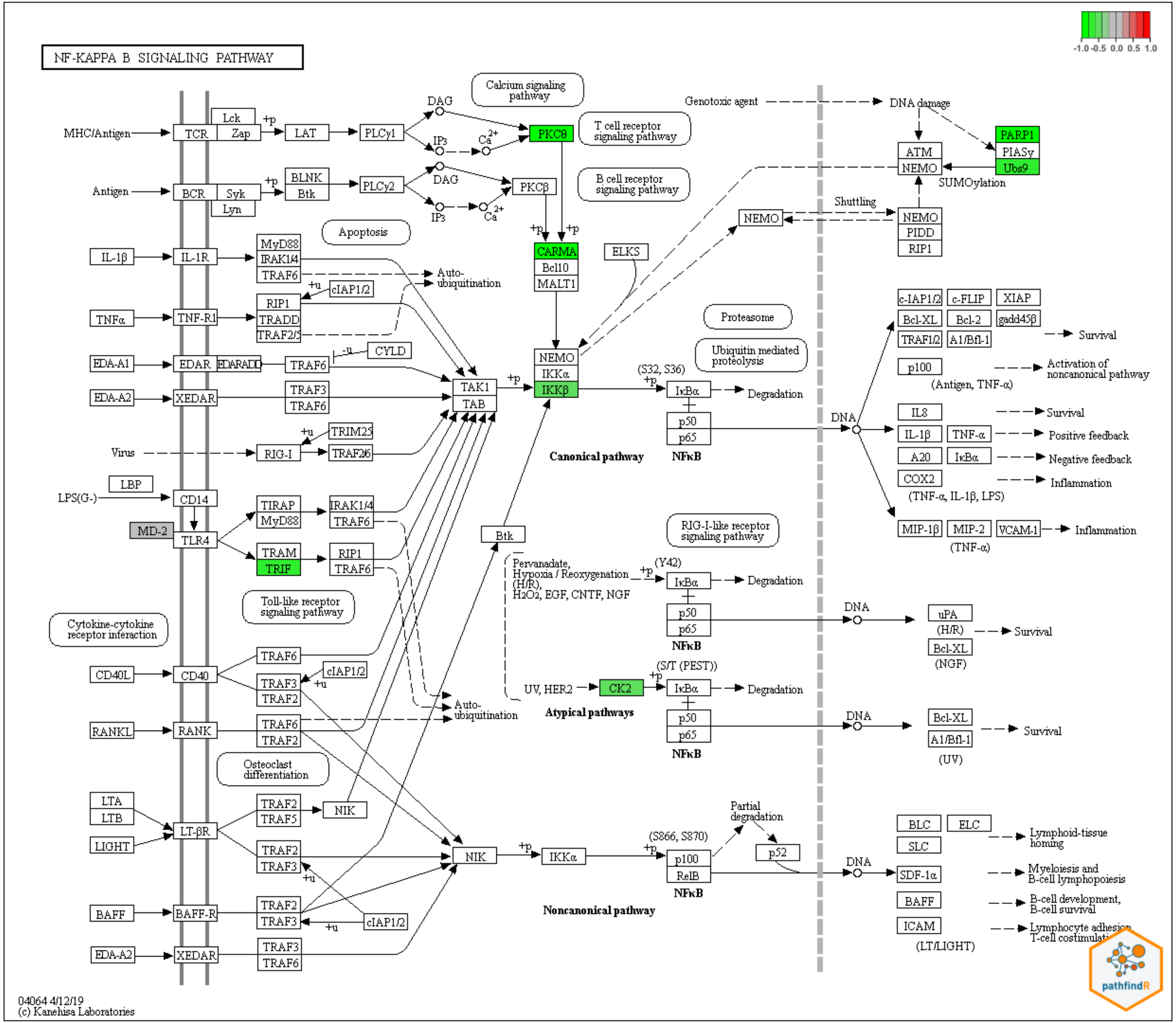
A table that lists all terms found to be enriched as well as lists of up- or down-regulated genes for each term. If it was requested, the term descriptions are linked to the visualizations of these terms, where affected color genes are colored by change values (if provided).

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).

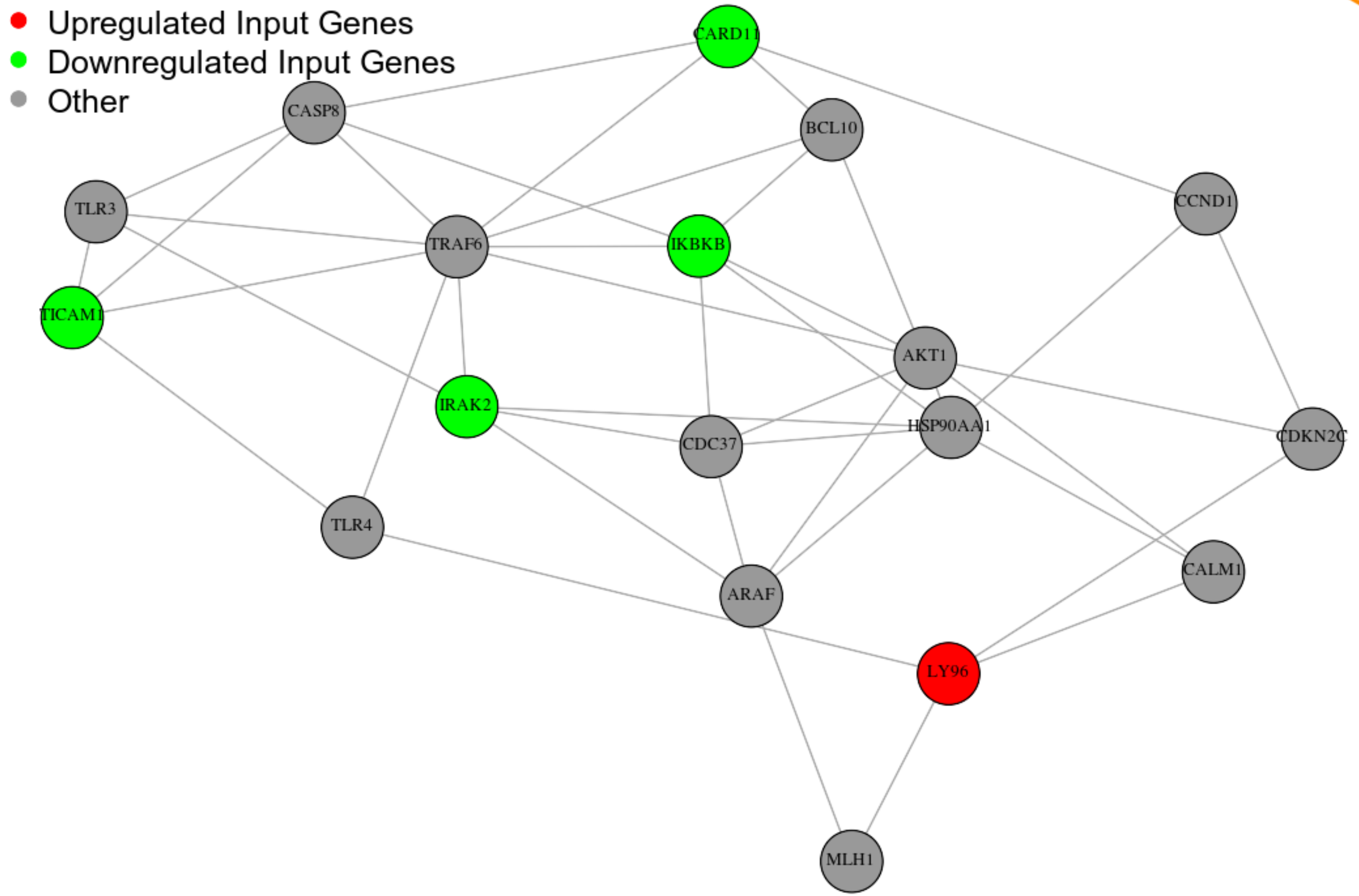
pathfindR - All Enriched Terms - KEGG

ID	Term_Description	Fold_Enrichment	occurrence	lowest_p	highest_p	Up_regulated	Down_regulated
hsa03040	Spliceosome	3.09750	1	1.1e-09	1.1e-09	SF3B6, LSM3, BUD31	SNRPB, SF3B2, U2AF2, PUF60, DDX23, EIF4A3, HNRNPA1, PCBP1, SRSF8, SRSF5
hsa00190	Oxidative phosphorylation	2.52397	1	2.9e-08	2.9e-08	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C, ATP6V1D, ATP6V0E1	ATP6V0E2
hsa05012	Parkinson disease	2.31877	1	4.9e-08	4.9e-08	NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	UBE2G1, VDAC1, SLC25A5
hsa04932	Non-alcoholic fatty liver disease (NAFLD)	1.98061	1	1.3e-07	1.3e-07	DDIT3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	IKBKB, FASLG
hsa03410	Base excision repair	4.80149	1	1.6e-07	1.6e-07	POLE4	MUTYH, APEX2, POLD2, PARP1
hsa05010	Alzheimer disease	2.26356	1	1.9e-07	1.9e-07	GAPDH, RTN3, NDUFA1, NDUFB3, UQCRQ, COX6A1, COX7A2, COX7C	CALM3, CALM1, ATP2A2
hsa04064	NF-kappa B signaling pathway	2.53519	1	3.0e-07	3.0e-07	LY96	PRKCQ, CARD11, TICAM1, IKBKB, UBE2I, CSNK2A2, PARP1

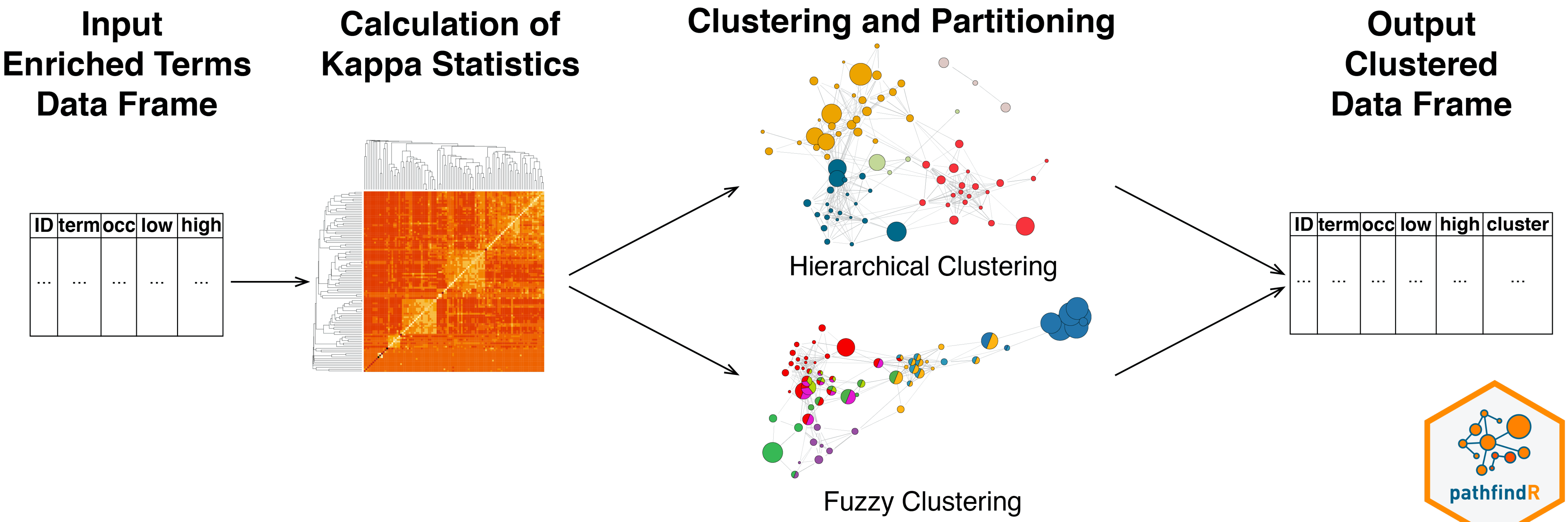


NF-KAPPA B SIGNALING PATHWAY
-1.0 -0.5 0.0 0.5 1.0
MHC/Antigen
TCR
Lck
Zap
+p
LAT
PLCγ1
DAG
IP3
Ca2+
PKCθ
Calcium signaling pathway
T cell receptor signaling pathway
Antigen
BCR
Syk
Lyn
BLNK
Btk
PLCγ2
PKCβ
B cell receptor signaling pathway
CARMA
Bcl10
MALT1
ELKS
IL-1β
IL-1R
MyD88
IRAK1/4
TRAF6
Auto-ubiquitination
cIAP1/2
+u
TNFα
TNF-R1
RIP1
TRADD
TRAF2/5
Apoptosis
EDA-A1
EDAR
EDARADD
TRAF6
-u
CYLD
EDA-A2
XEDAR
TRAF3
TRAF6
TAK1
TAB
NEMO
IKKα
IKKβ
(S32, S36)
IκBα
Degradation
p50
p65
NFκB
Canonical pathway
Proteasome
Ubiquitin mediated proteolysis
Virus
RIG-I
TRIM25
TRAF2/6
LBP
LPS(G-)
CD14
MD-2
TLR4
TIRAP
MyD88
IRAK1/4
TRAF6
TRAM
TRIF
RIP1
TRAF6
Toll-like receptor signaling pathway
Btk
Pervanadate, Hypoxia / Reoxygenation (H/R), H2O2, EGF, CNTF, NGF
(Y42)
RIG-I-like receptor signaling pathway
Cytokine-cytokine receptor interaction
CD40L
CD40
TRAF6
cIAP1/2
TRAF3
TRAF2
UV, HER2
CK2
(S/T (PEST))
Atypical pathways
RANKL
RANK
TRAF6
TRAF2
Osteoclast differentiation
LTA
LTB
LIGHT
LT-βR
TRAF2
TRAF5
NIK
TRAF2
TRAF3
NIK
IKKα
(S866, S870)
p100
RelB
Partial degradation
p52
Noncanonical pathway
BAFF
BAFF-R
TRAF2
TRAF3
cIAP1/2
EDA-A2
XEDAR
TRAF3
TRAF6
Genotoxic agent
DNA damage
ATM
NEMO
PARP1
PIASy
Ubc9
SUMOylation
NEMO
Shuttling
NEMO
PIDD
RIP1
c-IAP1/2
c-FLIP
XIAP
Bcl-XL
Bcl-2
gadd45β
TRAF1/2
A1/Bfl-1
Survival
p100
Activation of noncanonical pathway
(Antigen, TNF-α)
DNA
IL8
Survival
IL-1β
TNF-α
Positive feedback
A20
IκBα
Negative feedback
COX2
Inflammation
(TNF-α, IL-1β, LPS)
MIP-1β
MIP-2
VCAM-1
Inflammation
(TNF-α)
uPA
(H/R)
Survival
Bcl-XL
(NGF)
Bcl-XL
A1/Bfl-1
(UV)
BLC
ELC
SLC
Lymphoid-tissue homing
SDF-1α
Myeloiesis and B-cell lymphopoiesis
BAFF
B-cell development, B-cell survival
ICAM
Lymphocyte adhesion T-cell costimulati
(LT/LIGHT)
04064 4/12/19
(c) Kanehisa Laboratories
pathfindR

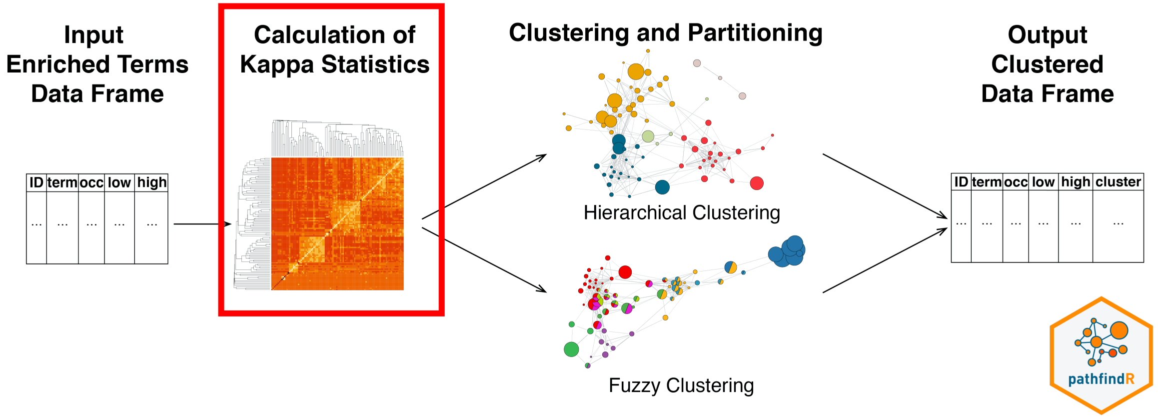
I-kappaB kinase-NF-kappaB signaling
Involved Gene Interactions in Biogrid



Clustering Workflow



pathfindR



Using
1 – kappa similarity
as distance metric for clustering

(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)

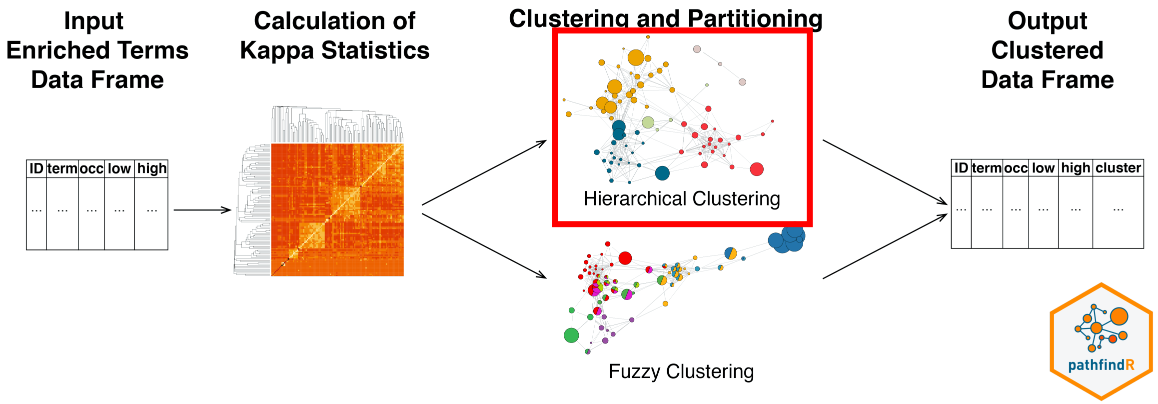
		Gene a: 1	Gene a: 0	Row total
Gene b	1	3 ($C_{1,1}$)	1 ($C_{0,1}$)	4 ($C_{1,\cdot}$)
	0	0 ($C_{0,1}$)	2 ($C_{0,0}$)	2 ($C_{0,\cdot}$)
Column total		3 ($C_{\cdot,1}$)	3 ($C_{\cdot,0}$)	6 (T_{ab})

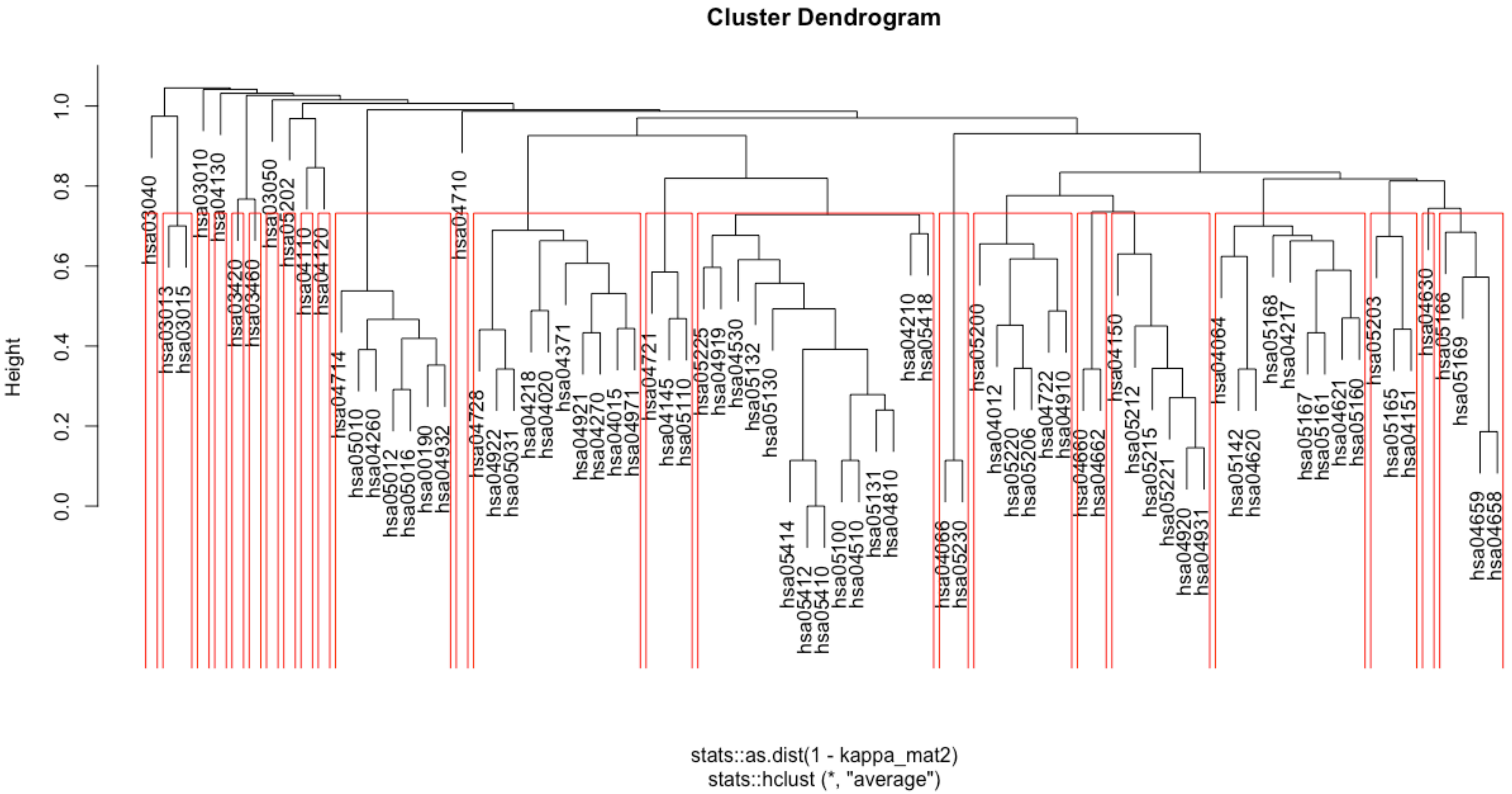
$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$

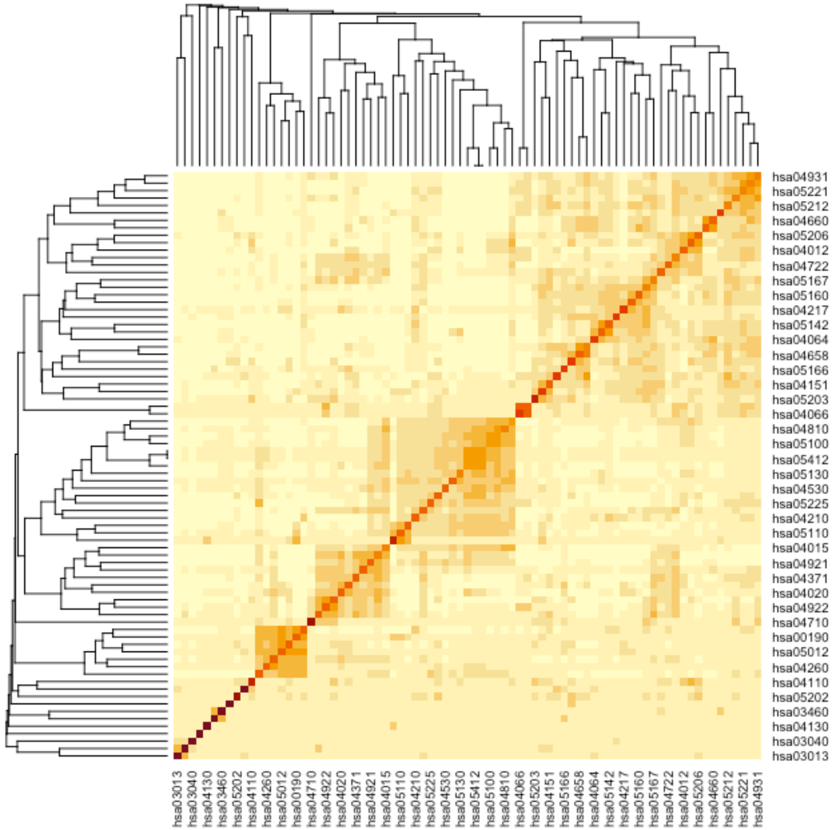
$$A_{ab} = \frac{C_{\cdot,1} \bullet C_{1,\cdot} + C_{\cdot,0} \bullet C_{0,\cdot}}{T_{ab} \bullet T_{ab}} = \frac{3 \bullet 4 + 3 \bullet 2}{6 \bullet 6} = 0.5$$

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.83 - 0.5}{1 - 0.5} = 0.66$$

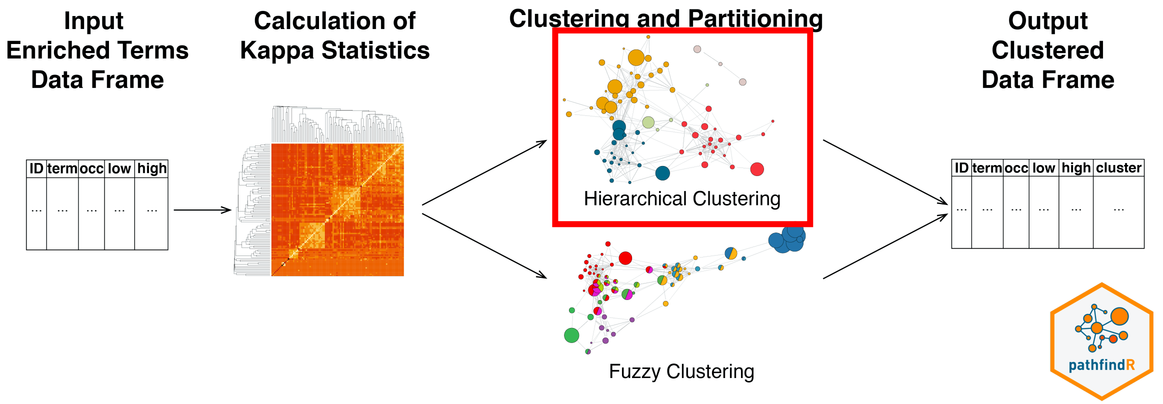
Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol. 2007;8(9):R183.



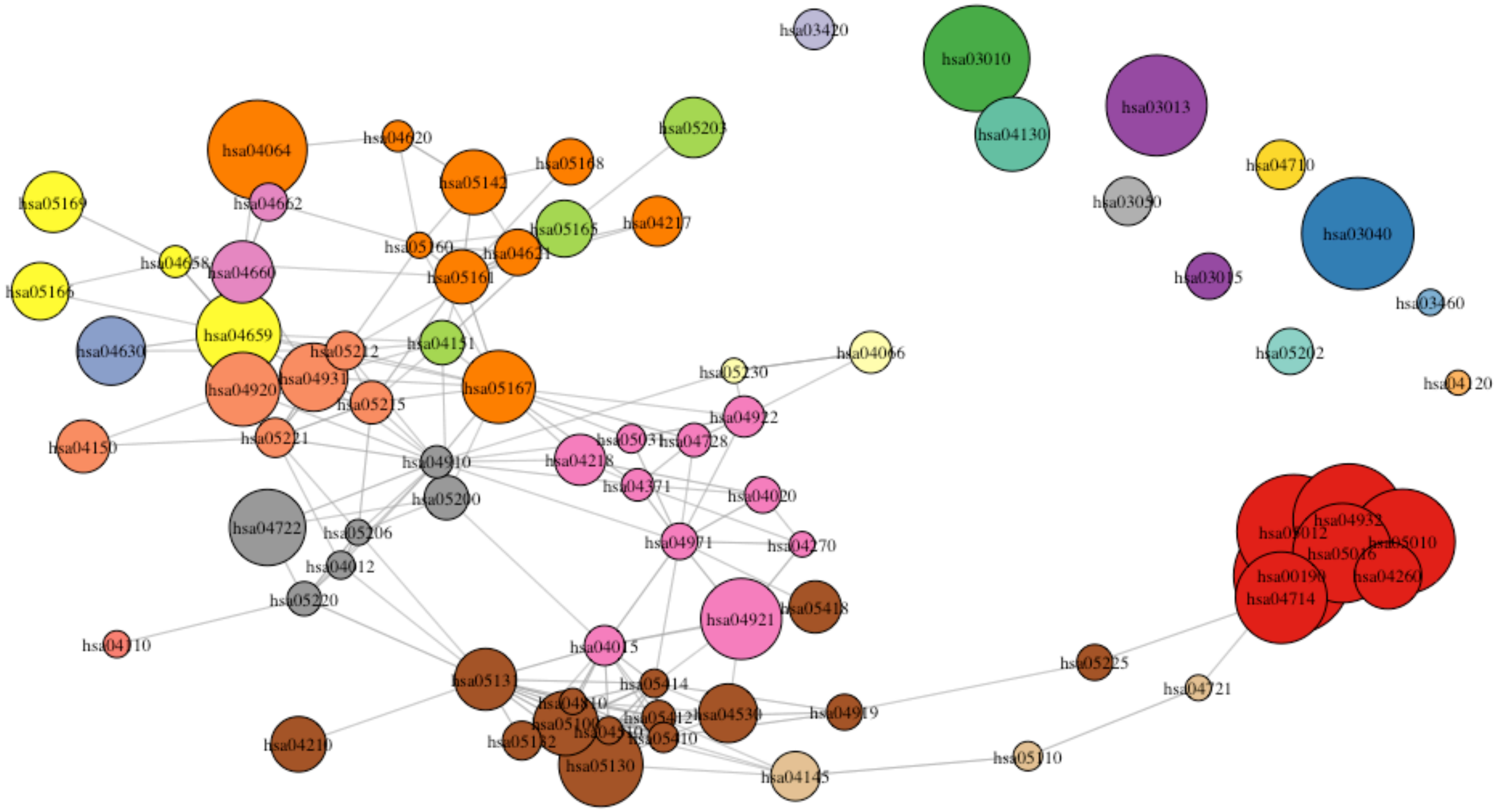


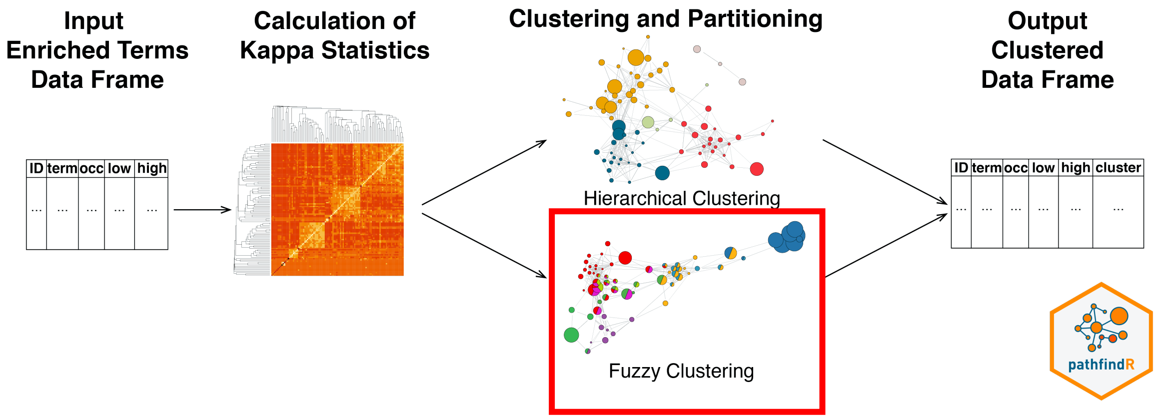


The optimal number of clusters is automatically determined by maximizing the average silhouette width

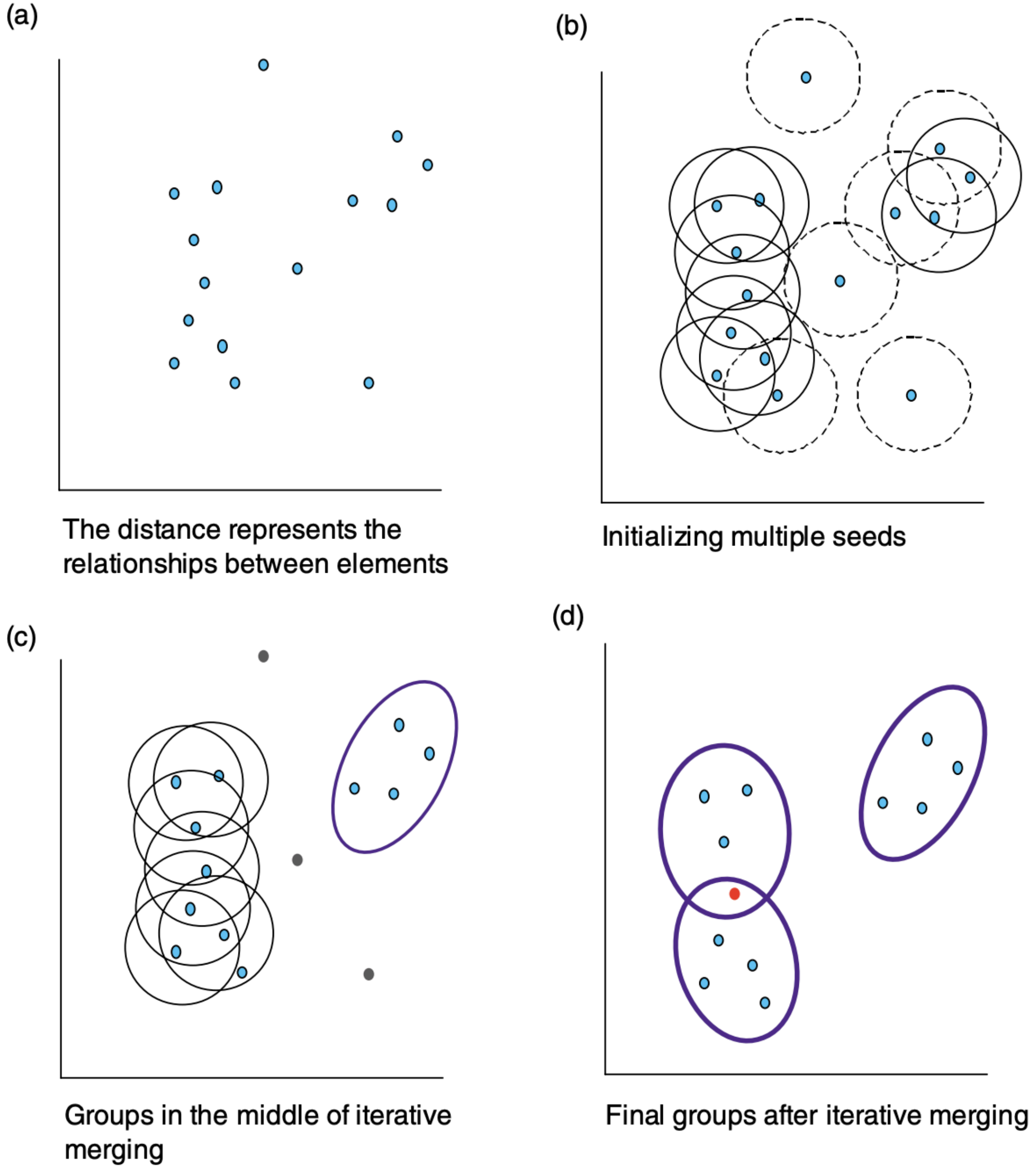


No links shown for
kappa < 0.35 (default)



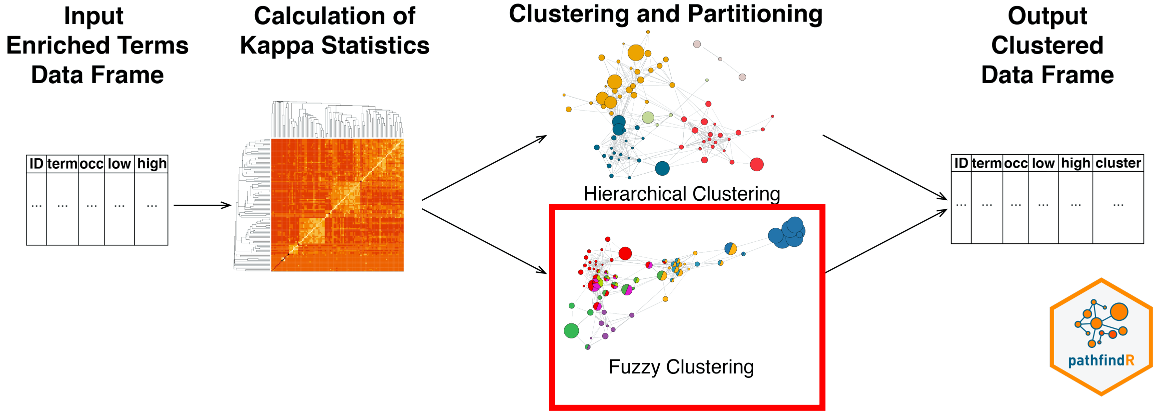


The heuristic fuzzy partition algorithm

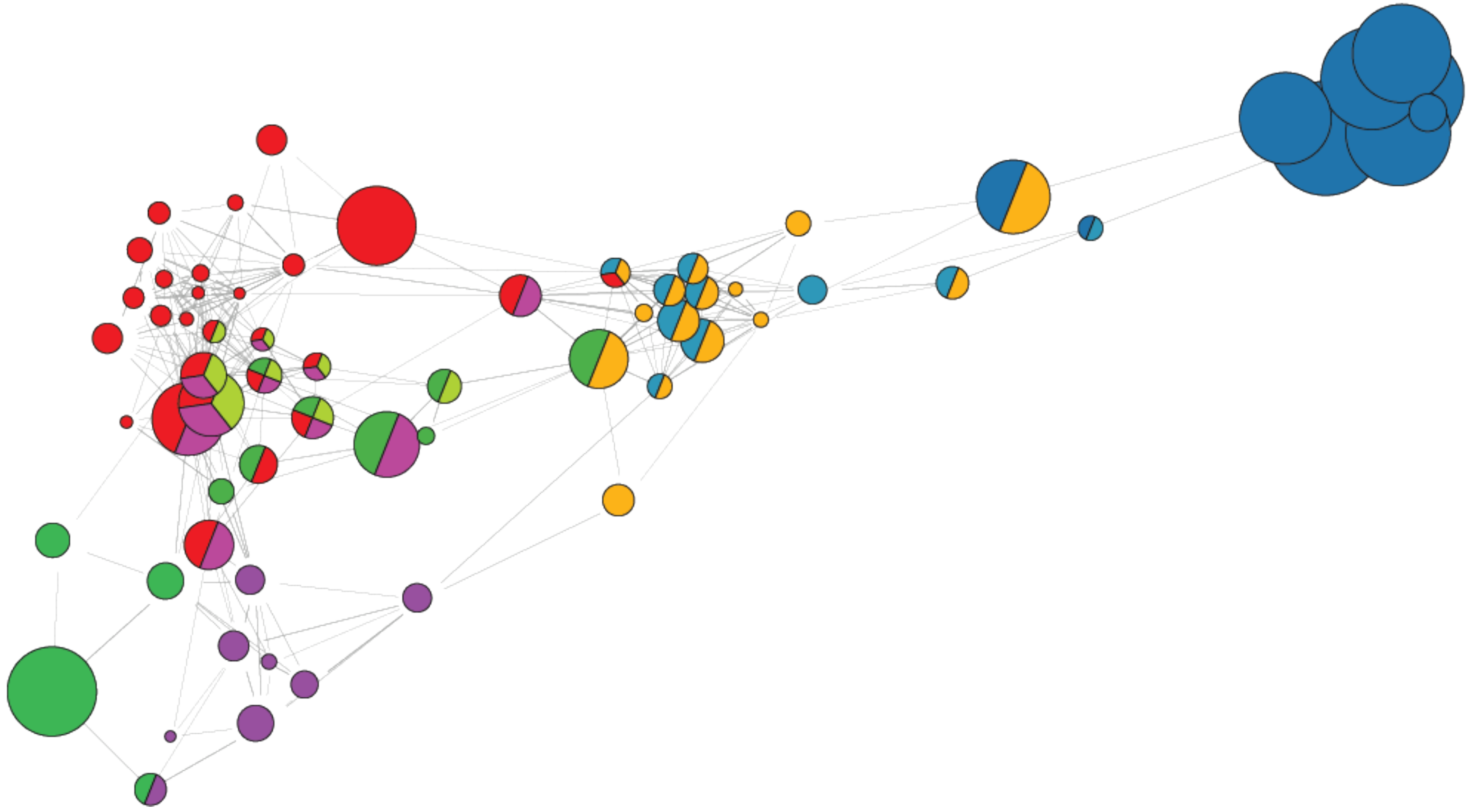


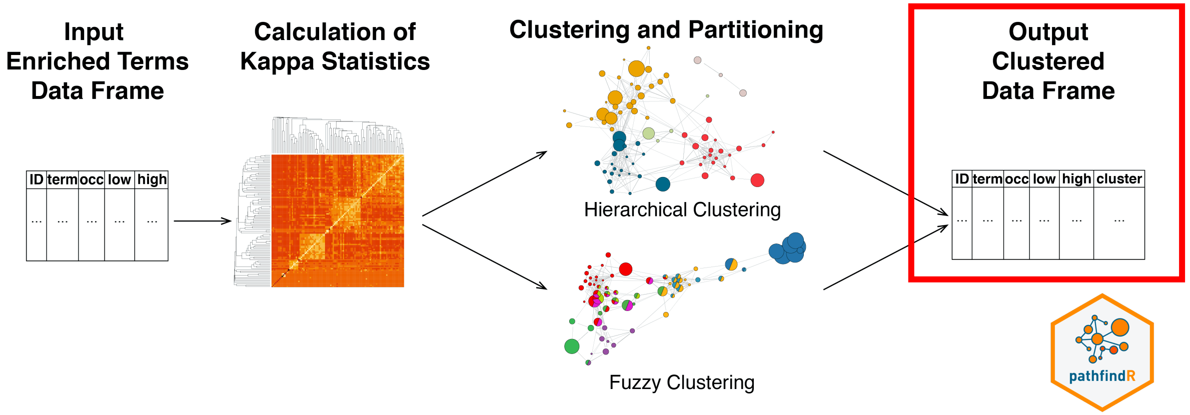
Using
1 – kappa similarity
as distance metric for clustering

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.



No links shown for
kappa < 0.35 (default)





Representative term selection

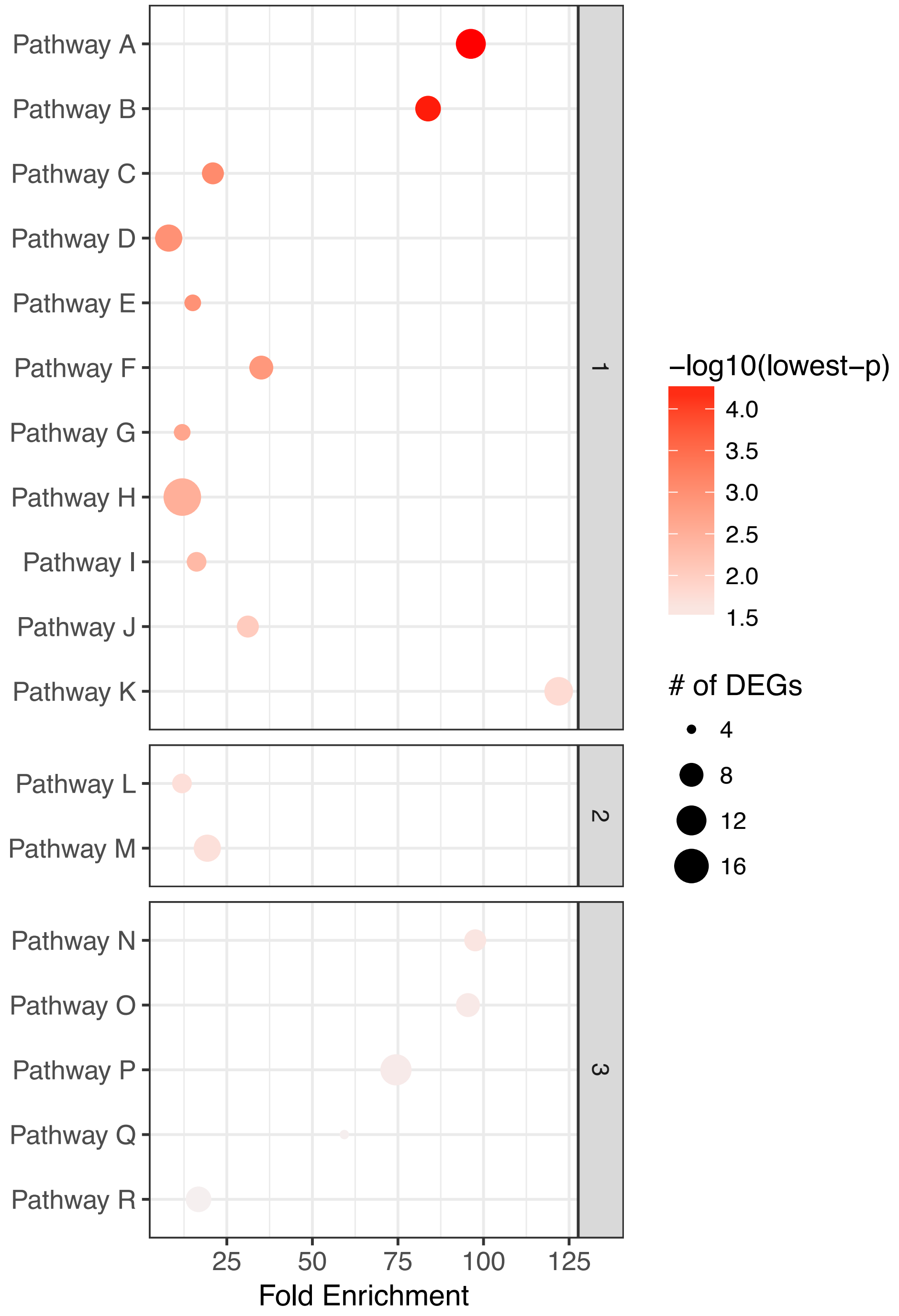
For each cluster, the representative term is chosen as the one with the lowest p value (default)

Note that this is an **ad hoc** decision and different approaches may be used:

- Highest fold enrichment
- The most biologically meaningful, etc.

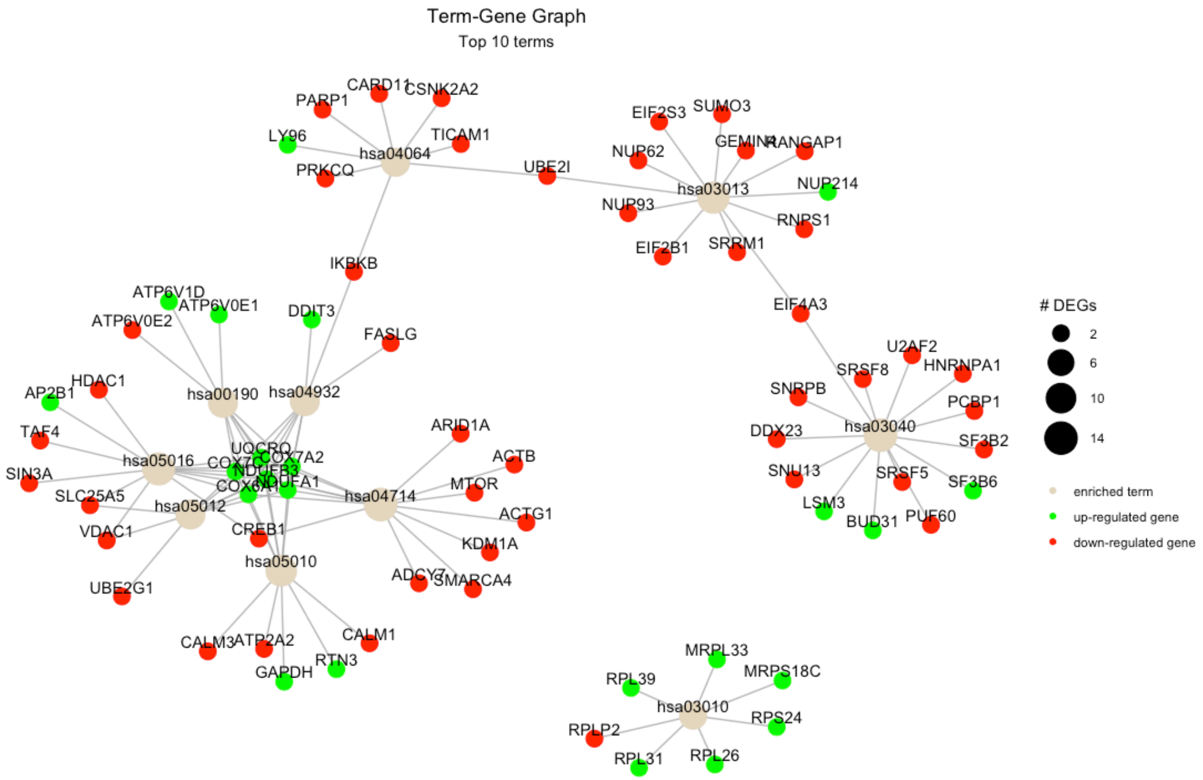
ID	Term_Description	Fold_Enrichment	occurrence	lowest_p	highest_p	Up_regulated	Down_regulated	Cluster	Status
hsa00190	Oxidative phosphorylation	71.863	10	2.61E-07	2.61E-07	NDUFB3, NDUFA1, COX7C	ATP6V0E2	1	Representative
hsa05012	Parkinson's disease	63.727	10	3.88E-07	3.88E-07	UQCRQ, COX6A1, COX7A2	VDAC1, 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

⋮

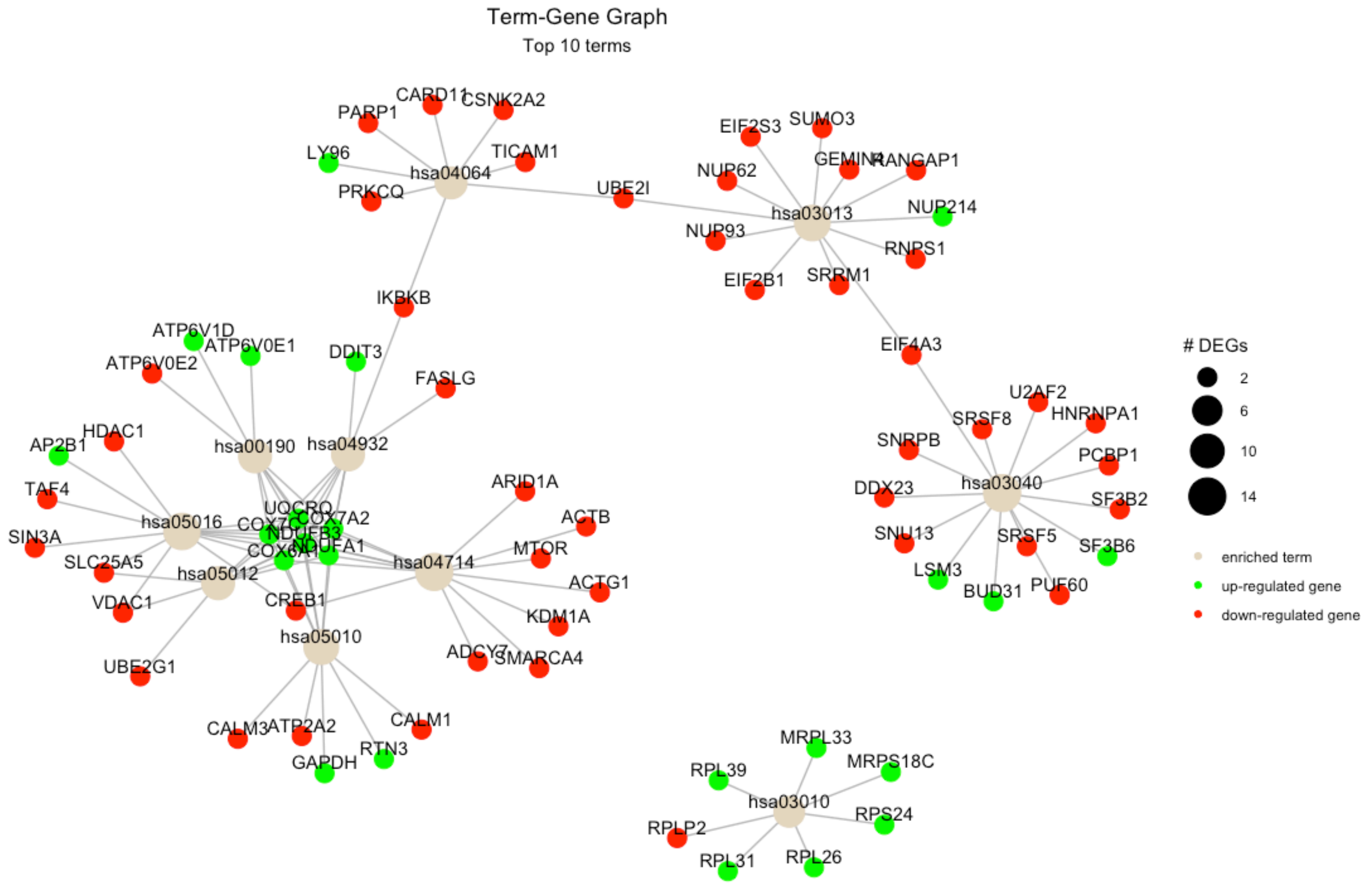

Pathway A
Pathway B
Pathway C
Pathway D
Pathway E
Pathway F
Pathway G
Pathway H
Pathway I
Pathway J
Pathway K
Pathway L
Pathway M
Pathway N
Pathway O
Pathway P
Pathway Q
Pathway R
1
2
3
25
50
75
100
125
Fold Enrichment
−log10(lowest−p)
4.0
3.5
3.0
2.5
2.0
1.5
# of DEGs
4
8
12
16

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
hsa04064
hsa03013
hsa00190
hsa04932
hsa05016
hsa05012
hsa04714
hsa05010
hsa03040
hsa03010
CARD11
CSNK2A2
PARP1
LY96
TICAM1
PRKCQ
UBE2I
EIF2S3
SUMO3
GEMIN4
RANGAP1
NUP62
NUP214
NUP93
RNPS1
EIF2B1
SRRM1
IKBKB
EIF4A3
ATP6V1D
ATP6V0E1
ATP6V0E2
DDIT3
FASLG
HDAC1
AP2B1
TAF4
SIN3A
SLC25A5
VDAC1
UBE2G1
UQCRQ
COX7A2
COX7C
NDUFB3
COX6A1
NDUFA1
CREB1
ARID1A
ACTB
MTOR
ACTG1
KDM1A
ADCY7
SMARCA4
CALM3
ATP2A2
CALM1
RTN3
GAPDH
U2AF2
SRSF8
HNRNPA1
SNRPB
PCBP1
DDX23
SF3B2
SNU13
SRSF5
SF3B6
LSM3
BUD31
PUF60
MRPL33
RPL39
MRPS18C
RPS24
RPLP2
RPL31
RPL26
DEGs
2
6
10
14
enriched term
up-regulated gene
down-regulated gene

Agglomerated Scoring of Terms per Subject

Conceptual Background

For an experiment matrix (containing expression, methylation, etc. values), the rows of which are genes and the columns of which are samples, we denote:

- E as a matrix of size $m \times n$
- G as the set of all genes in the experiment $G = E_{i\cdot}, \ \ i \in [1, m]$
- S as the set of all samples in the experiment $S = E_{j\cdot}, \ \ \in [1, n]$

We next define the gene score matrix GS (the standardized experiment matrix, also of size $m \times n$) as:

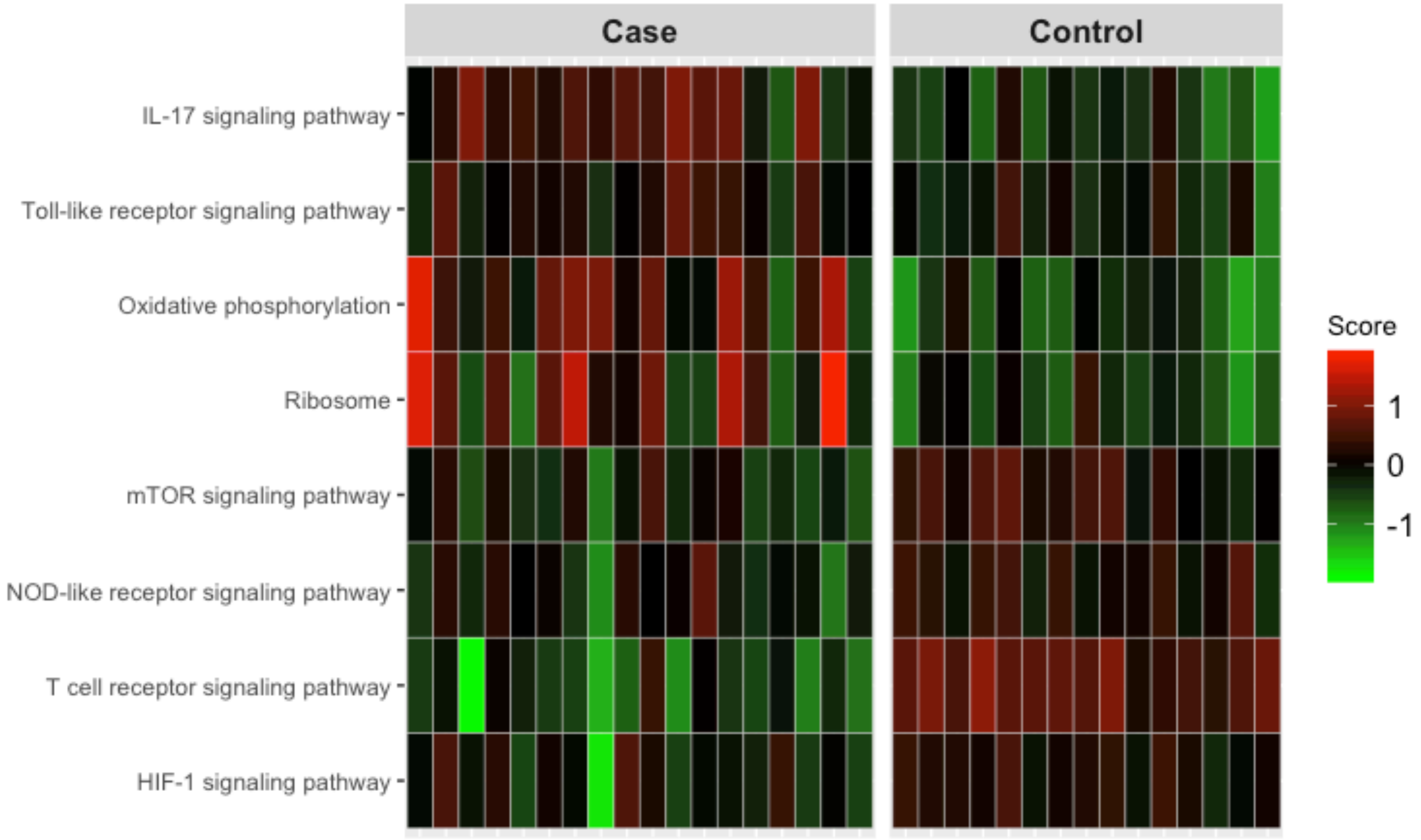
$GS_{gs} = \frac{E_{gs} - \bar{e_g}}{s_g}$

where $g \in G$, $s \in S$, $\bar{e_g}$ is the mean of all values for gene g and $\bar{s_g}$ is the standard deviation of all values for gene g.

We next denote T to be a set of terms (where each $t \in T$ is a set of term-related genes, i.e., $t = \{g_x, ..., g_y\} \subset G$) and finally define the agglomerated term scores matrix TS (where rows correspond to genes and columns corresponds to samples s.t. the matrix has size $|T| \times n$) as:

$TS_{ts} = \frac{1}{|t|} \sum_{g \in t} GS_{gs}$, where $t \in T$ and $s \in S$.

Heatmap of Agglomerated Scores
grouped by Case/Control



Installation (CRAN release version – latest 1.4.0)

Installation – Bioconductor Dependencies

```
if (!requireNamespace("BiocManager", quietly = TRUE))
        install.packages("BiocManager")
BiocManager::install("KEGGREST")
BiocManager::install("KEGGgraph")
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

Installation (Development version)

From GitHub

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

or from DockerHub

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

pathfindR

CRAN release 1.4.0

pathfindR: Enrichment Analysis Utilizing Active Subnetworks

Enrichment analysis enables researchers to uncover mechanisms underlying a phenotype. However, conventional methods for enrichment analysis do not take into account protein-protein interaction information, resulting in incomplete conclusions. pathfindR is a tool for enrichment analysis utilizing active subnetworks. The main function identifies active subnetworks in a protein-protein interaction network using a user-provided list of genes and associated p values. It then performs enrichment analyses on the identified subnetworks, identifying enriched terms (i.e. pathways or, more broadly, gene sets) that possibly underlie the phenotype of interest. pathfindR also offers functionalities to cluster the enriched terms and identify representative terms in each cluster, to score the enriched terms per sample and to visualize analysis results. The enrichment, clustering and other methods implemented in pathfindR are described in detail in Ulgen E, Ozisik O, Sezerman OU. 2019. pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks. Front. Genet. <doi:10.3389/fgene.2019.00858>.

Version:	1.4.0
Depends:	R (≥ 3.6)
Imports:	DBI, AnnotationDbi, doParallel, foreach, rmarkdown, org.Hs.eg.db, ggplot2, ggraph, fpc, grDevices, igraph, R.utils, magick, KEGGREST, KEGGgraph, knitr
Suggests:	testthat (≥ 2.1.0), covr
Published:	2019-11-08
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:	Introduction to pathfindR Step-by-Step Execution of the pathfindR Enrichment Workflow pathfindR Analysis for non-Homo-sapiens organisms
Package source:	pathfindR_1.4.0.tar.gz
Windows binaries:	r-devel: pathfindR_1.4.0.zip, r-release: pathfindR_1.4.0.zip, r-oldrel: pathfindR_1.3.0.zip
OS X binaries:	r-release: not available, r-oldrel: not available
Old sources:	pathfindR archive

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