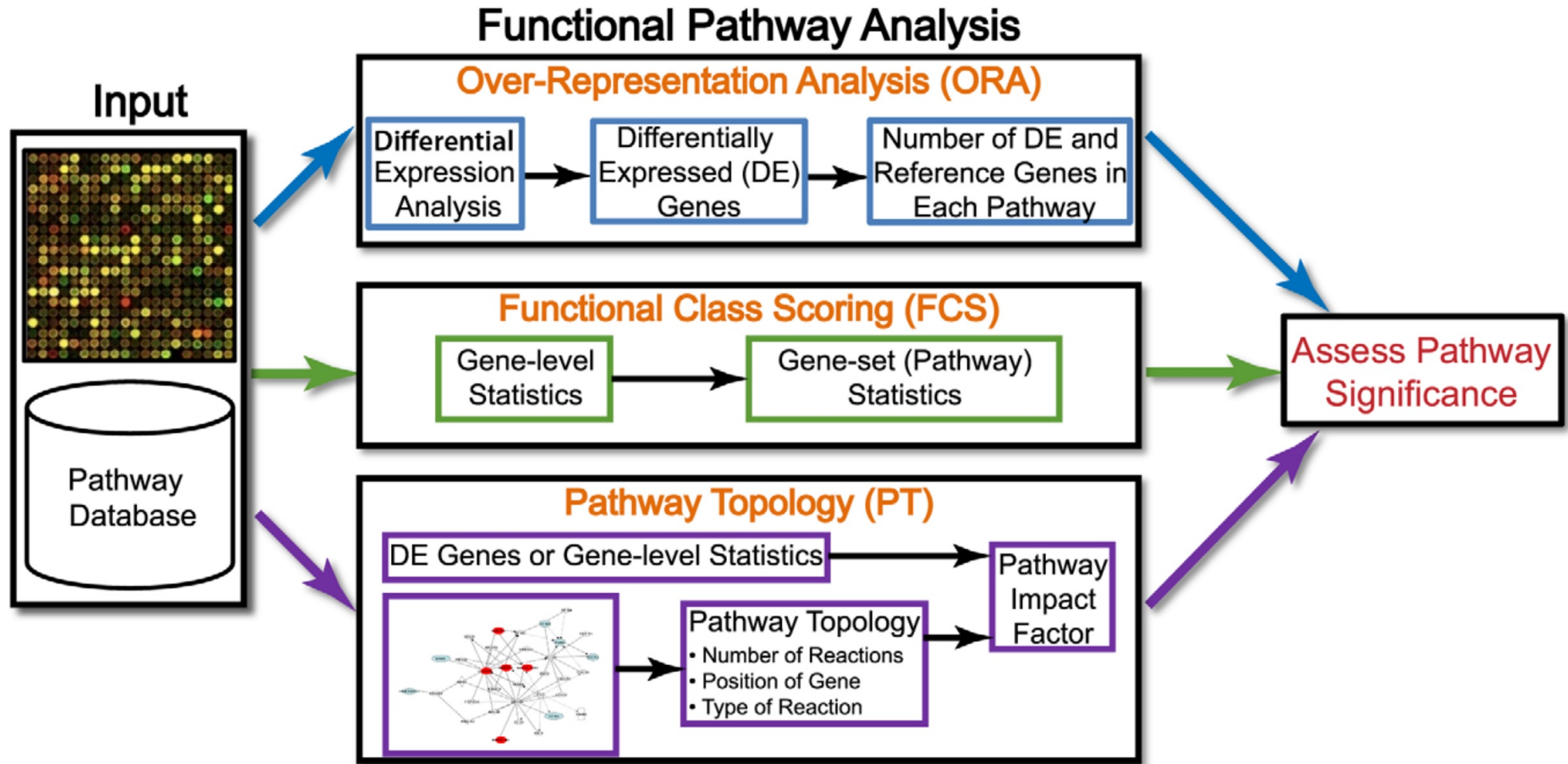# New in Version 2.4

- implemented a new color_kegg_pathway() function using ggkegg to create colored KEGG pathway ggplot objects (instead of using `KEGGREST` to obtain the colored PNG files, which no longer works)

- renamed the visualize_hsa_KEGG() function to visualize_KEGG_diagram() to reflect this is now able to handle KEGG pathway enrichment results from any organism

- updated the visualize_terms(), visualize_term_interactions() and visualize_KEGG_diagram() functions so that they now return a list of ggplot objects (named by term ID)

- updated the get_kegg_gsets() function to also use ggkegg for fetching genes per pathway data

- updated the get_biogrid_pin() function so that it can now determine the latest version and download/process it from BioGRID

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

Functional Pathway Analysis

Input

Over-Representation Analysis (ORA)
Differential Expression Analysis → Differentially Expressed (DE) Genes → Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)
Gene-level Statistics → Gene-set (Pathway) Statistics

Pathway Topology (PT)
DE Genes or Gene-level Statistics → Pathway Topology • Number of Reactions • Position of Gene • Type of Reaction → Pathway Impact Factor

Pathway Database

Assess Pathway Significance

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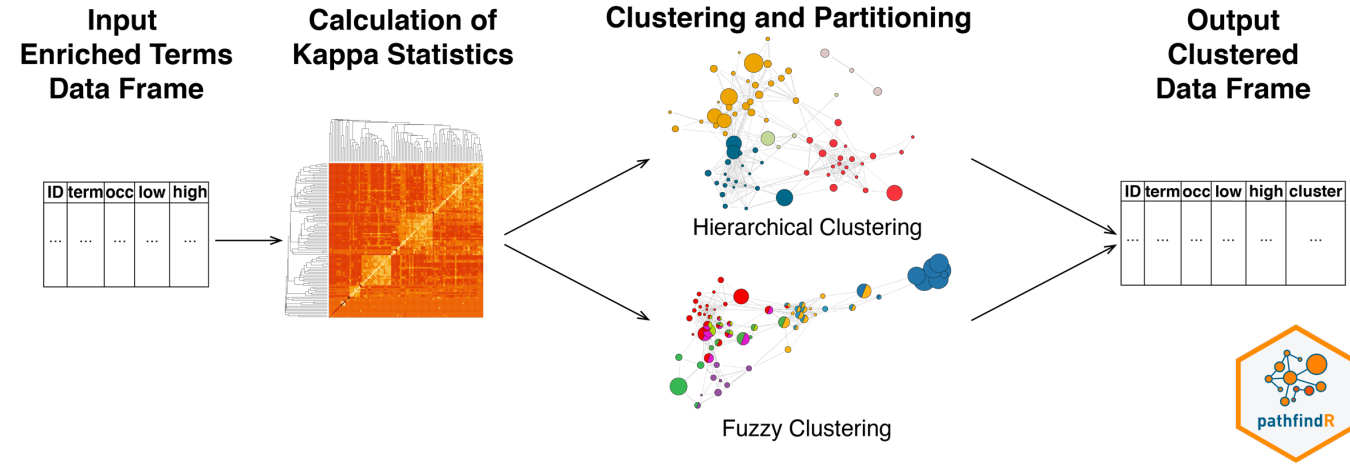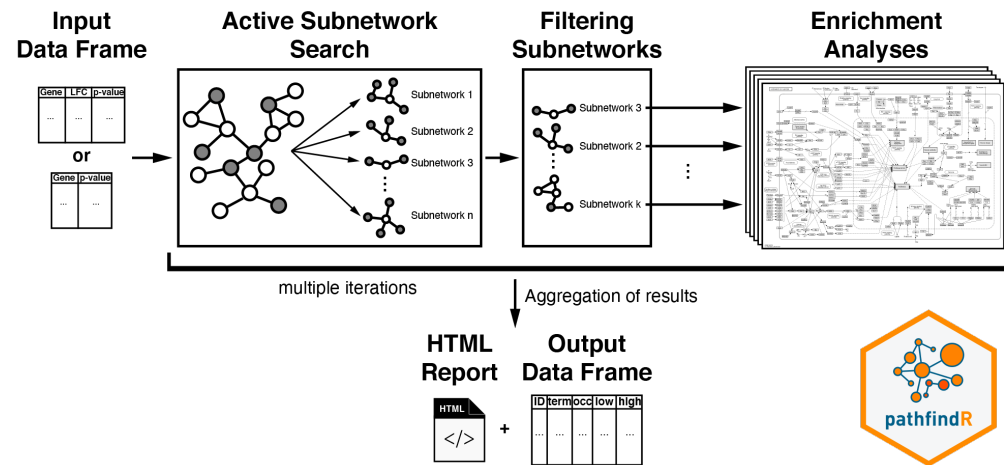
pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks

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
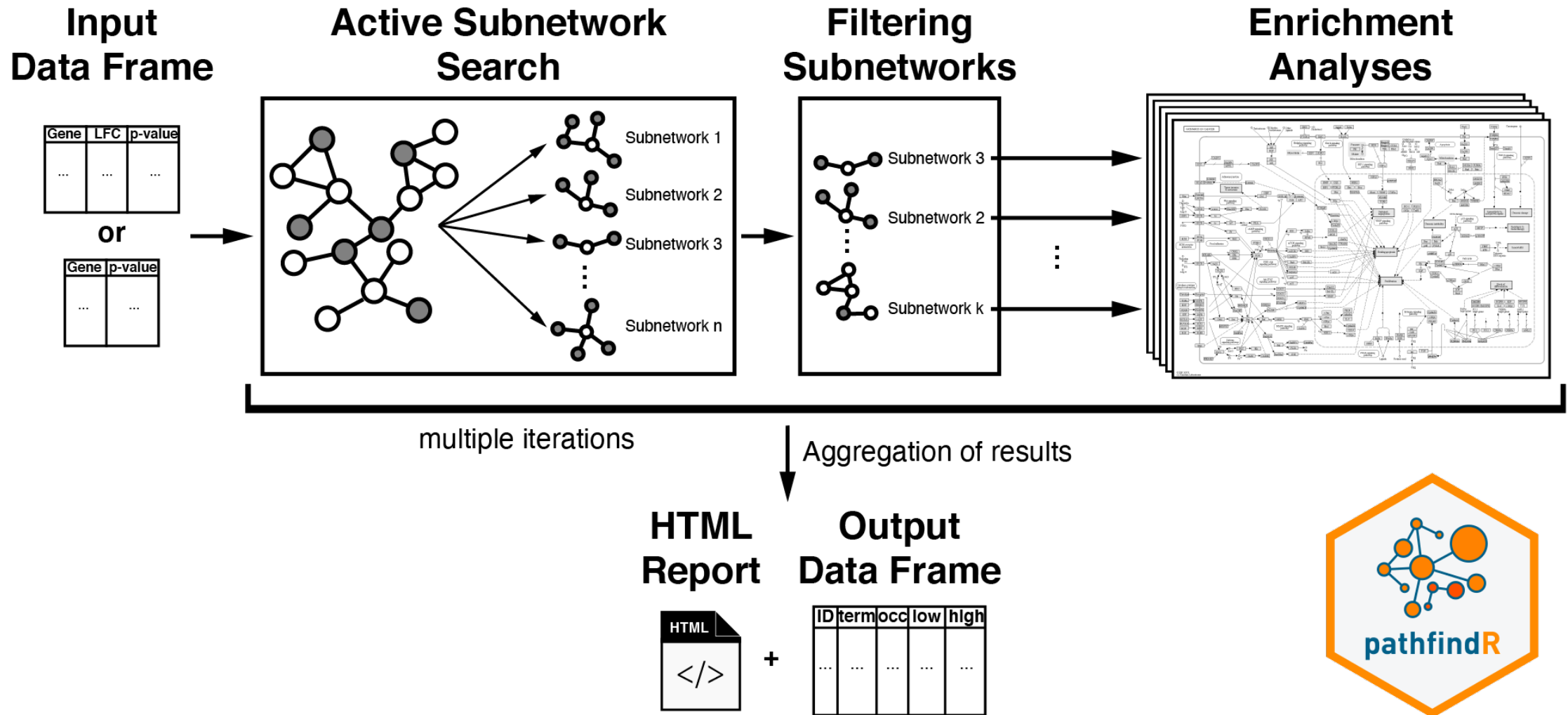[2]Department of Computer Engineering, Electrical & Electronics Faculty, Yildiz Technical University, Istanbul, Turkey
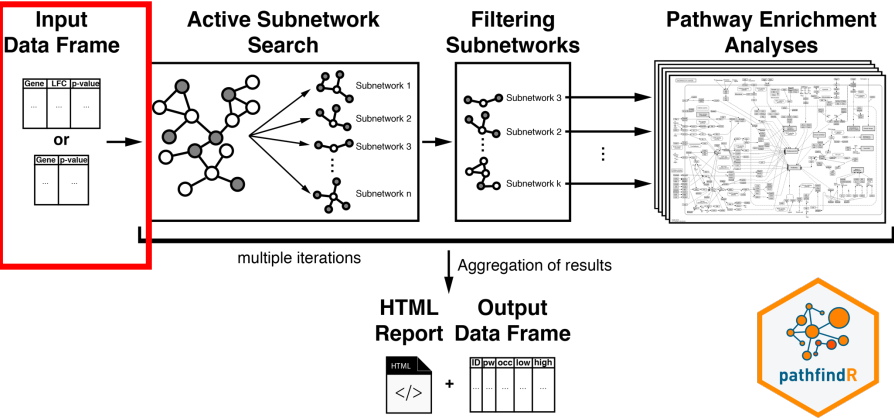
- Using input genes, pathfindR identifies sets of genes that form active subnetworks within a protein-protein interaction network

  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

  - Combine and compare 2 pathfindR enrichment results

  - Create various visualizations of the analysis
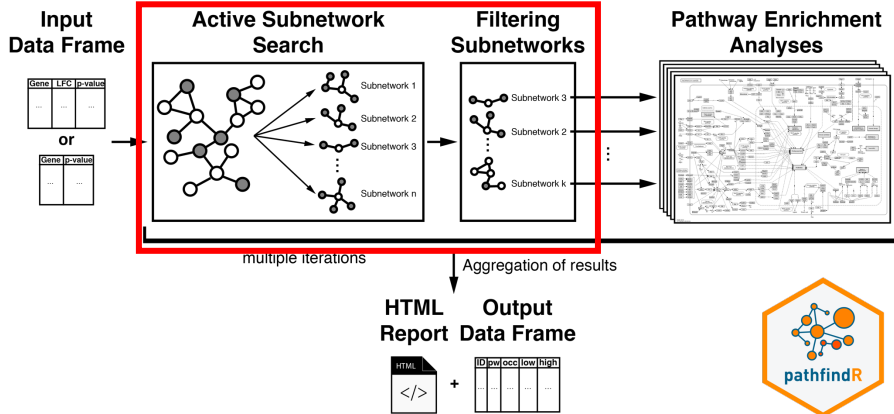
Active Subnetwork-oriented Enrichment Workflow

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

Input Data Frame

Gene	LFC	p-value

or

Gene	p-value

Active Subnetwork Search — Subnetwork 1, Subnetwork 2, Subnetwork 3, ... Subnetwork n

Filtering Subnetworks — Subnetwork 3, Subnetwork 2, ... Subnetwork k

Pathway Enrichment Analyses

multiple iterations

Aggregation of results

HTML Report **Output Data Frame**

ID	pw	occ	low	high

pathfindR

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.

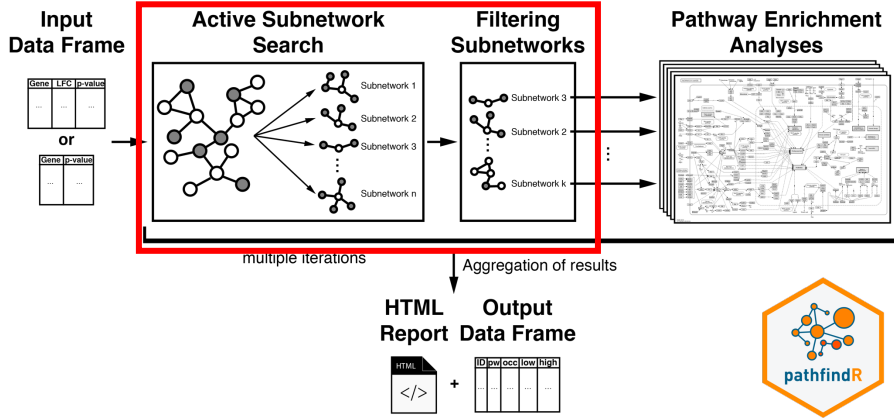$$z_i = \Phi^{-1}(1 - p_i) \qquad (1)$$

$$z_A = \frac{1}{\sqrt{k}} \sum_{i \, A} z_i \qquad (2)$$

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).

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

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

- *The user may also use get_pin_file() for obtaining organism-specific PIN data*

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

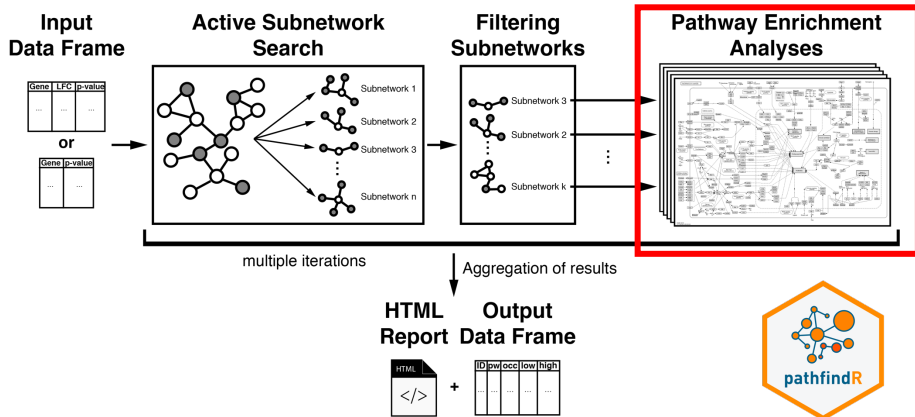*default options for pathfindR*

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).
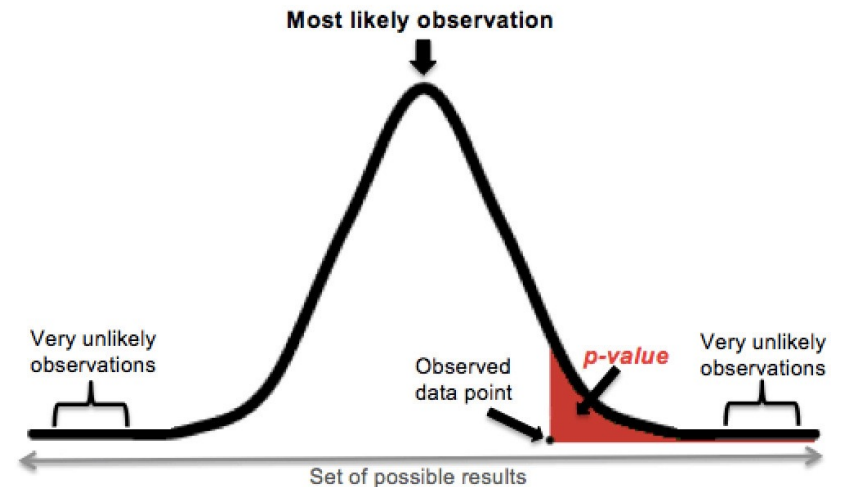
Choice of Active Subnetwork Search Method

- 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
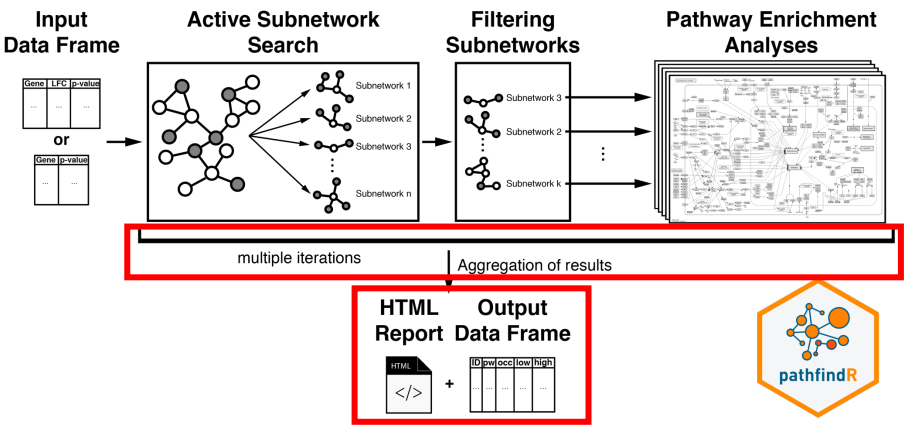
One-sided Hypergeometric Testing

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

- 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

  - *The user may also use get_gene_sets_list() for obtaining organism-specific gene sets list from KEGG, Reactome and MSigDB*
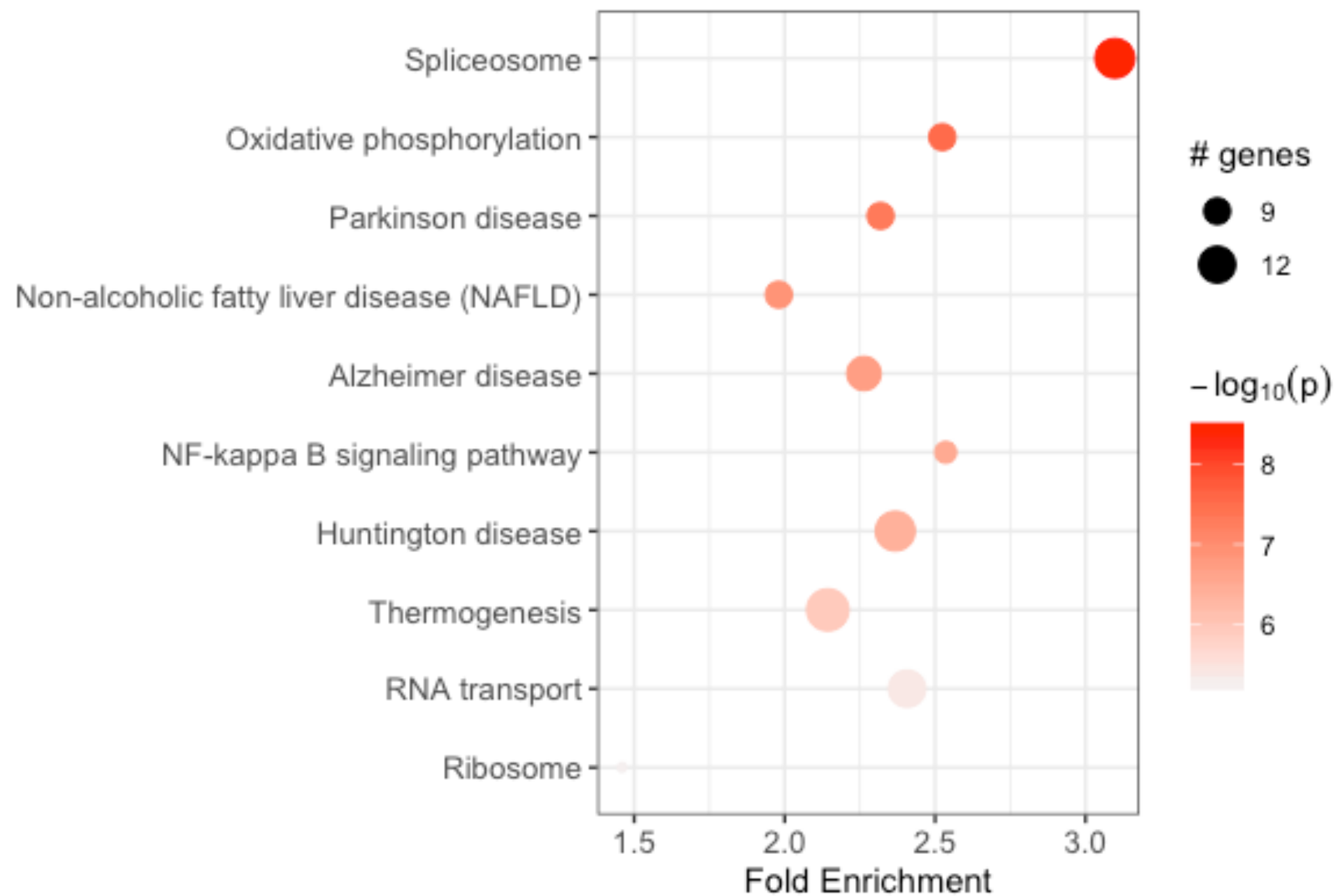

Most likely observation

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

*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

⋮

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).
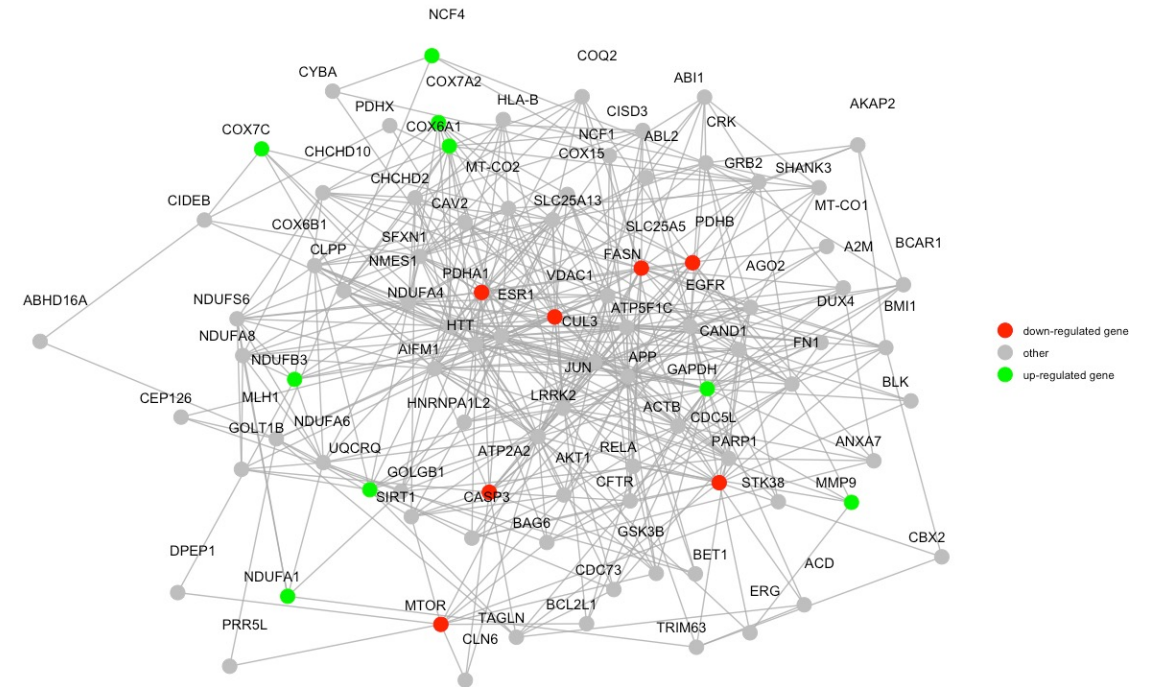
pathfindR - All Enriched Terms - KEGG

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

Visualize Term Diagram

visualize_terms()

- is_KEGG_result=TRUE (default):
  - Visualization of KEGG pathway diagram, colored by "change" values

- is_KEGG_result=FALSE:
  - Visualization of interactions of genes involved in the given enriched term

Enriched Term Clustering

**Input
Enriched Terms
Data Frame**

**Calculation of
Kappa Statistics**

Clustering and Partitioning

**Output
Clustered
Data Frame**

ID	term	occ	low	high
...	...	...	...	...

Hierarchical Clustering

Fuzzy Clustering

ID	term	occ	low	high	cluster
...	...	...	...	...	...

pathfindR

Input
Enriched Terms
Data Frame

ID	term	occ	low	high
...	...	...	...	...

Calculation of
Kappa Statistics

Clustering and Partitioning

Hierarchical Clustering

Fuzzy Clustering

pathfindR

Output
Clustered
Data Frame

ID	term	occ	low	high	cluster
...	...	...	...	...	...

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	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})

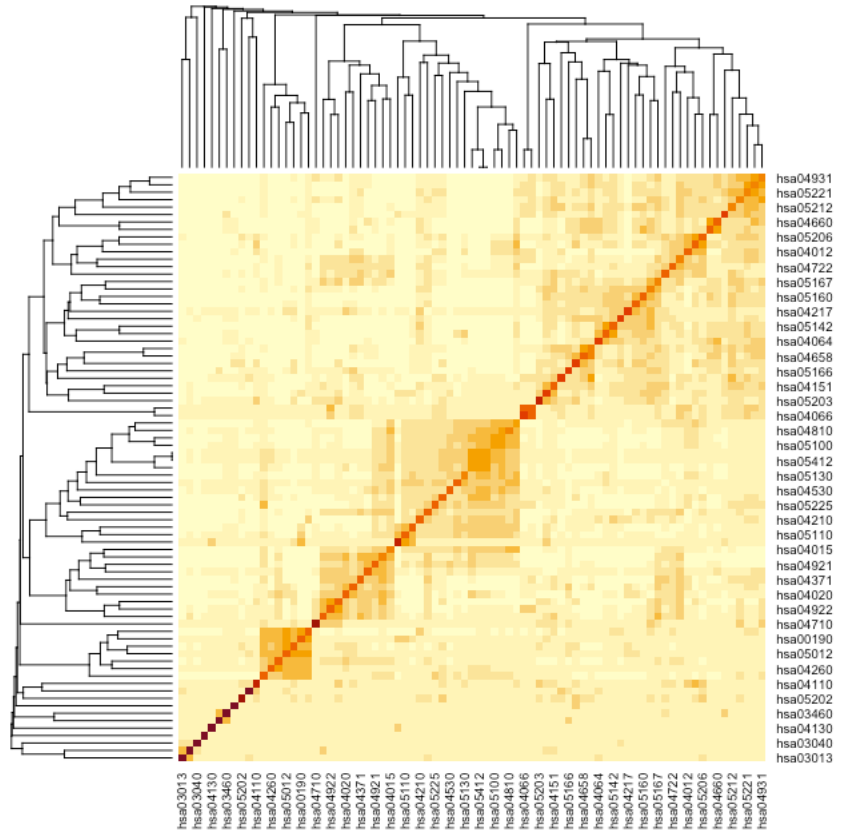$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$

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

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

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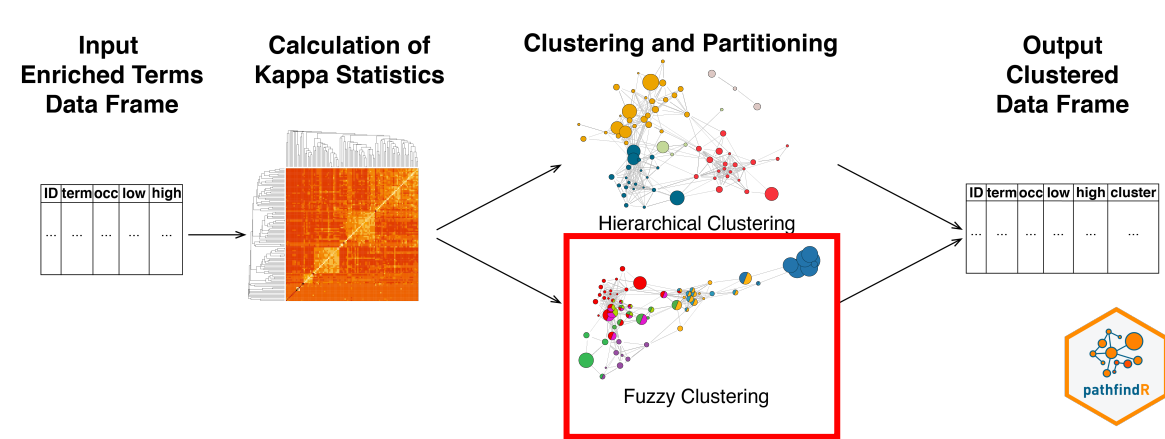
The optimal number of clusters is automatically determined by maximizing the average silhouette width

No links shown for kappa < 0.35 (default)

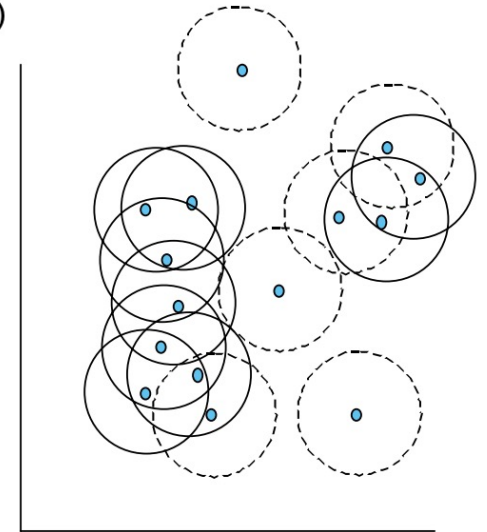# The heuristic fuzzy partition algorithm

Input Enriched Terms Data Frame

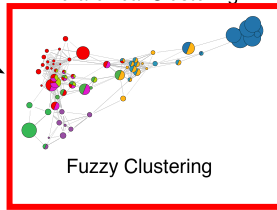| ID	term	occ	low	high
...	...	...	...	...

Calculation of Kappa Statistics

Clustering and Partitioning

Hierarchical Clustering

Fuzzy Clustering

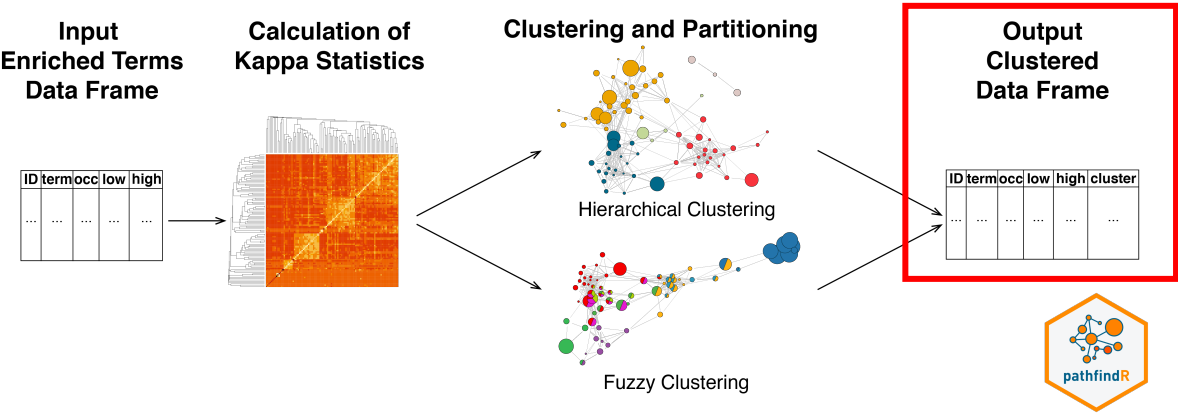**Output Clustered Data Frame**

ID	term	occ	low	high	cluster
...	...	...	...	...	...

pathfind**R**

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.

(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

Input
Enriched Terms
Data Frame

ID	term	occ	low	high
...	...	...	...	...
...	...	...	...	...

Calculation of
Kappa Statistics

Clustering and Partitioning

Hierarchical Clustering

Fuzzy Clustering

Output
Clustered
Data Frame

ID	term	occ	low	high	cluster
...	...	...	...	...	...
...	...	...	...	...	...

pathfind**R**

No links shown for
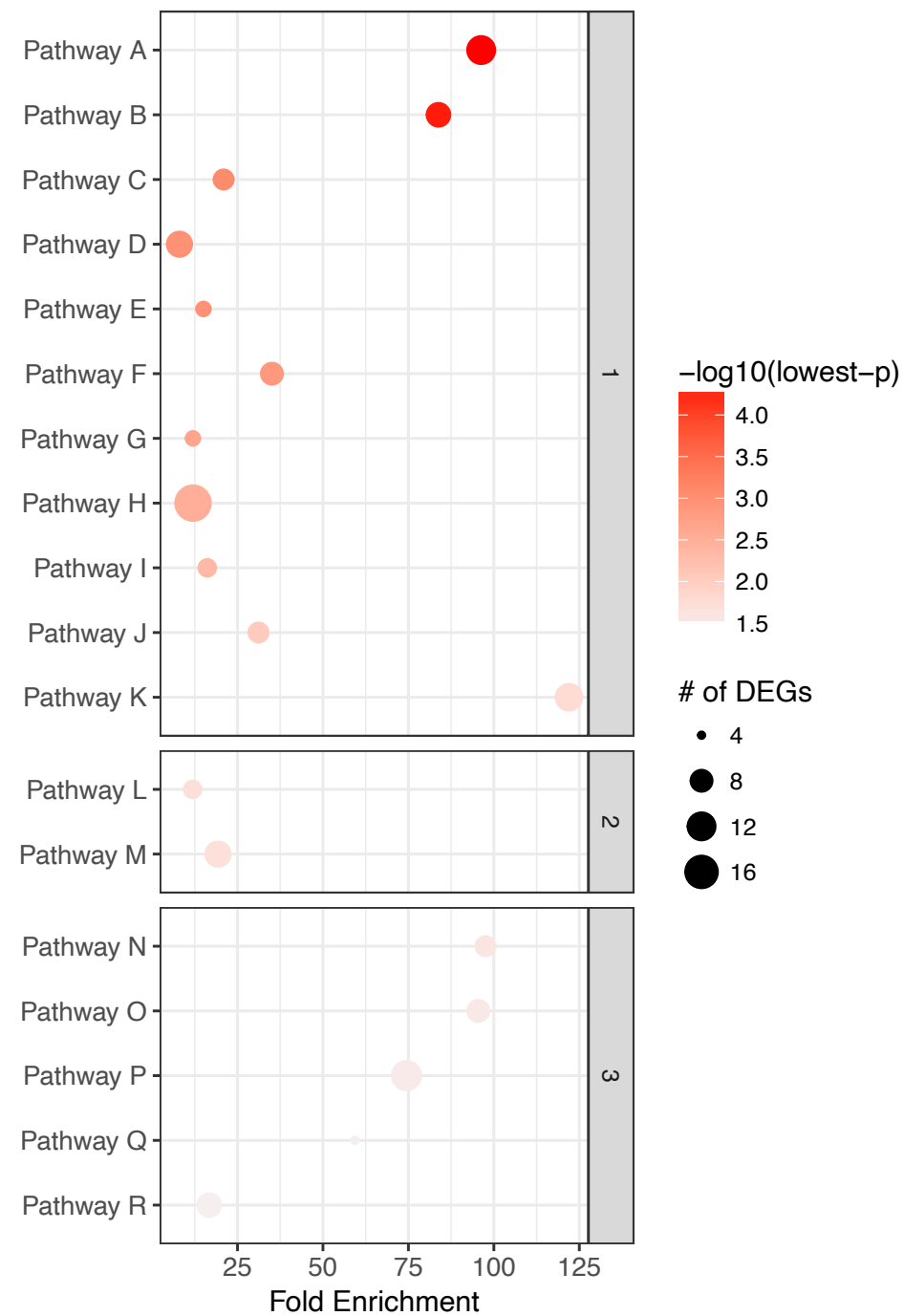kappa < 0.35 (default)

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

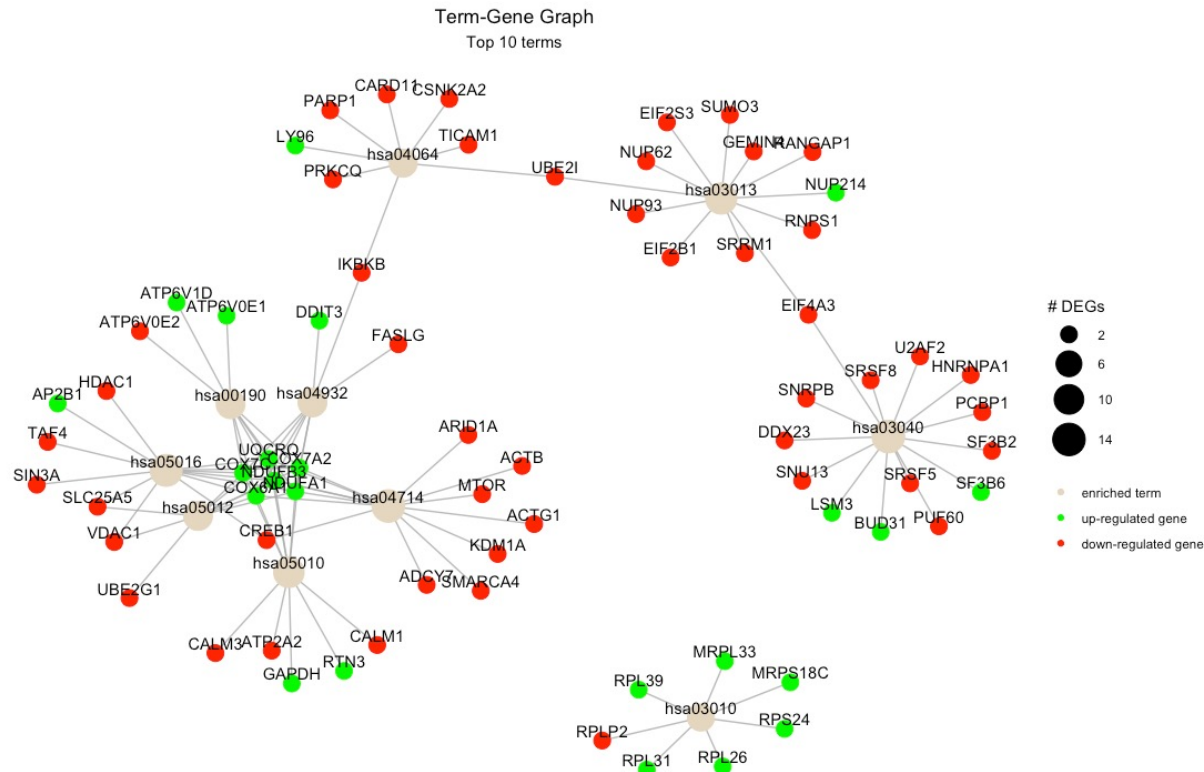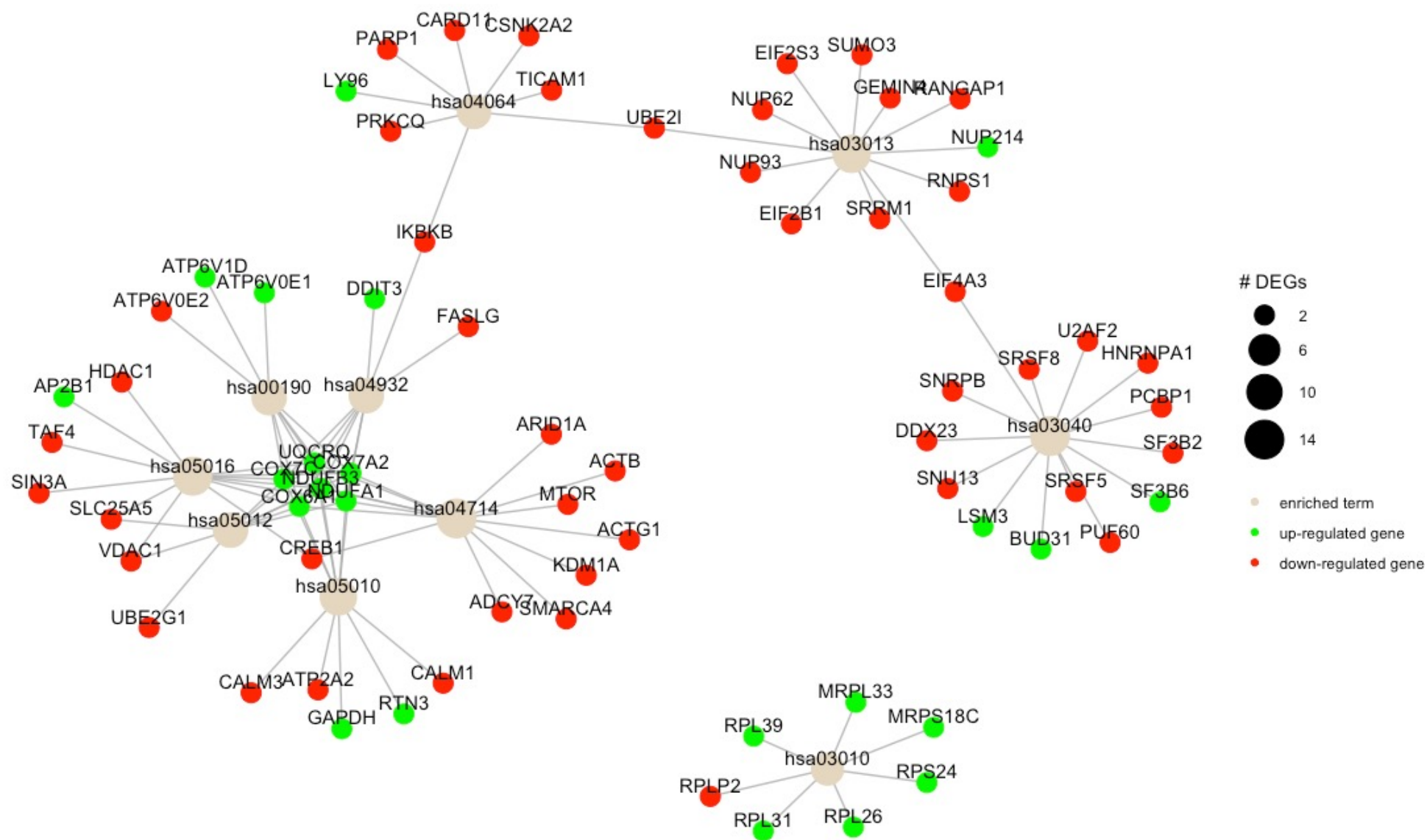⋮

Term-Gene Graph

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:
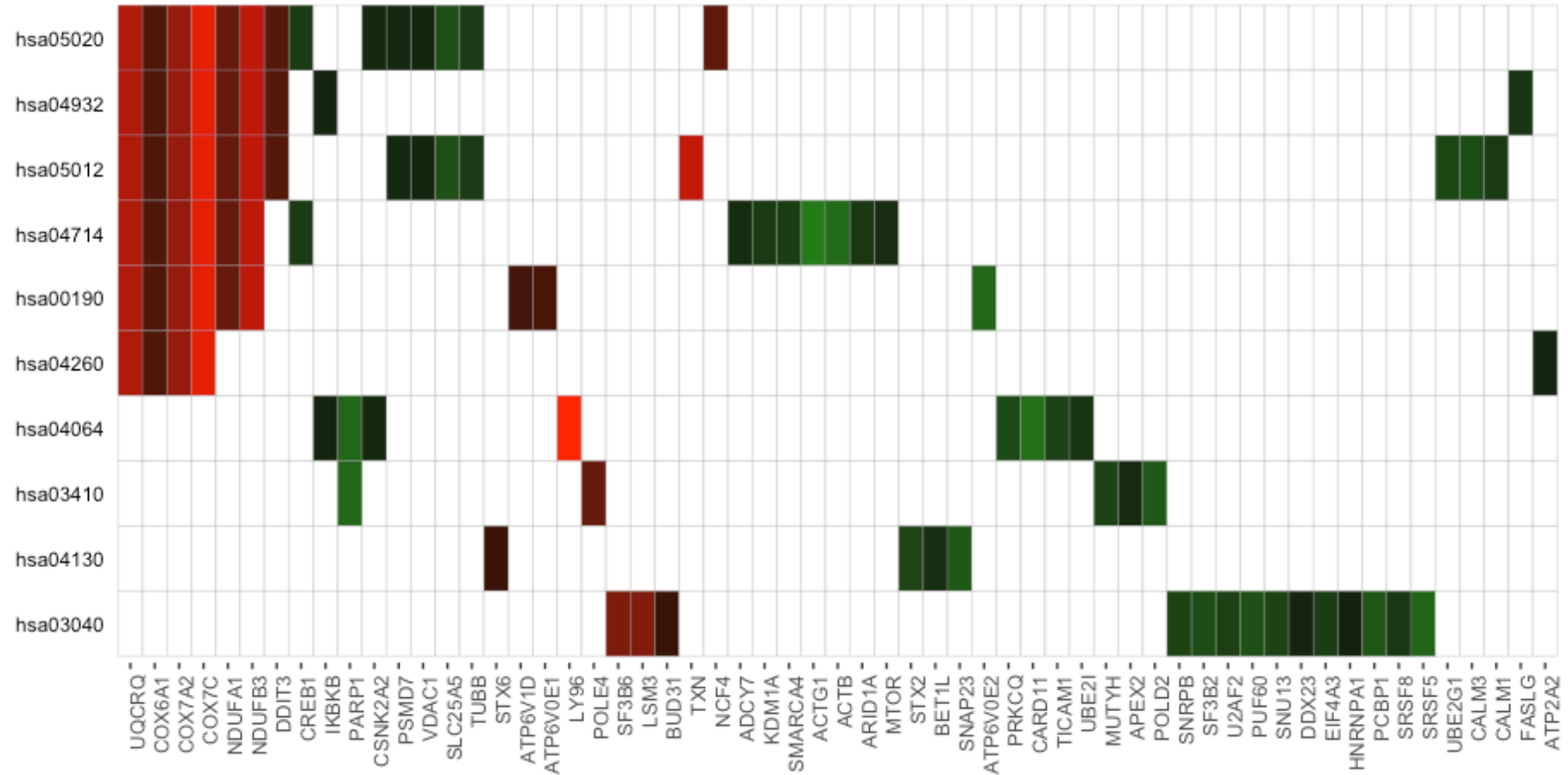 - the number of genes (default)
  - the $-\log_{10}$(p value)
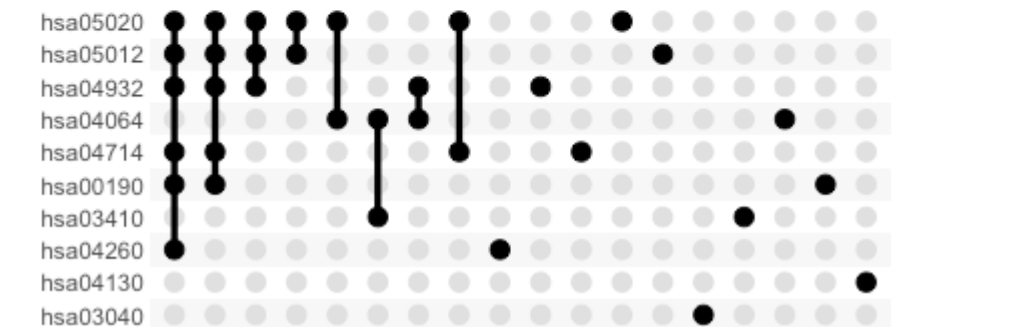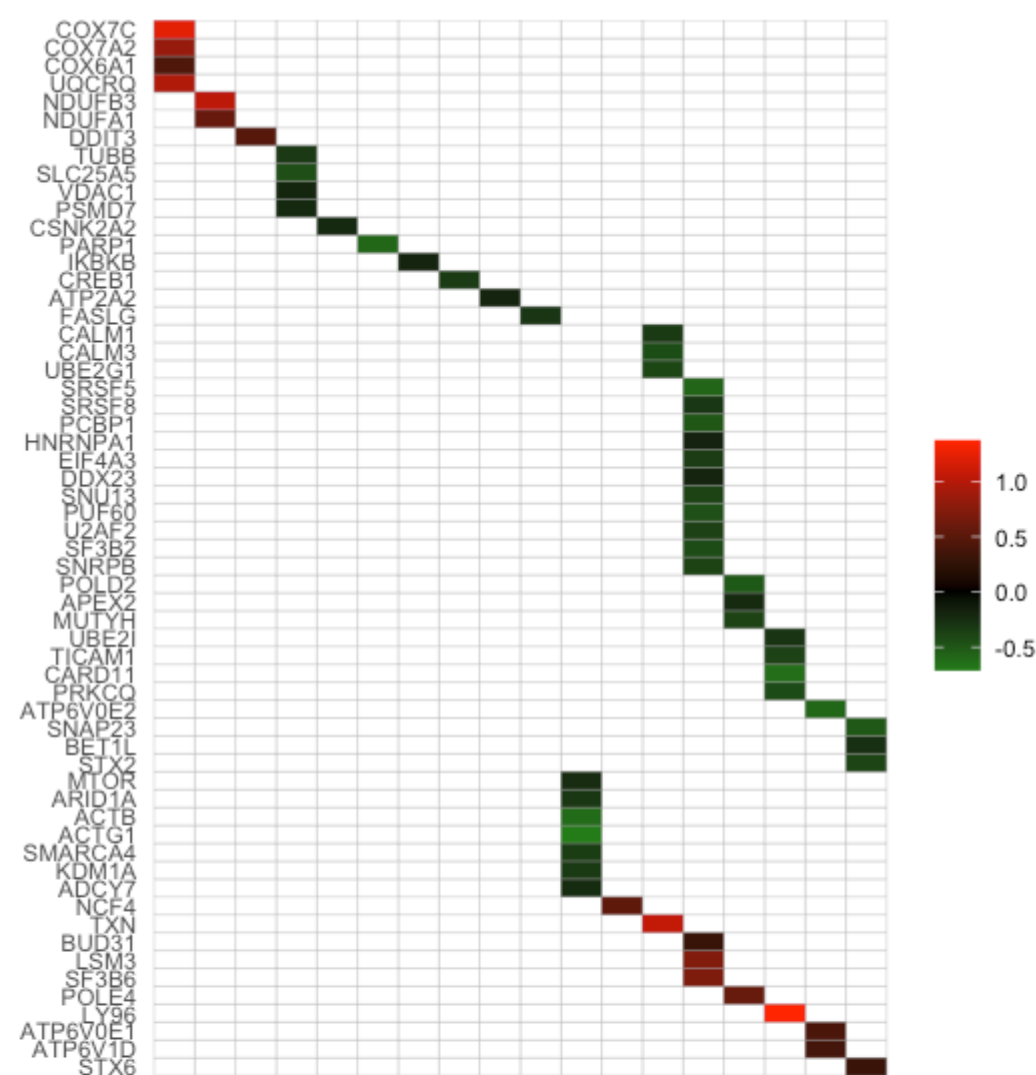
Term-Gene Graph
Top 10 terms

Term – Gene Heatmap

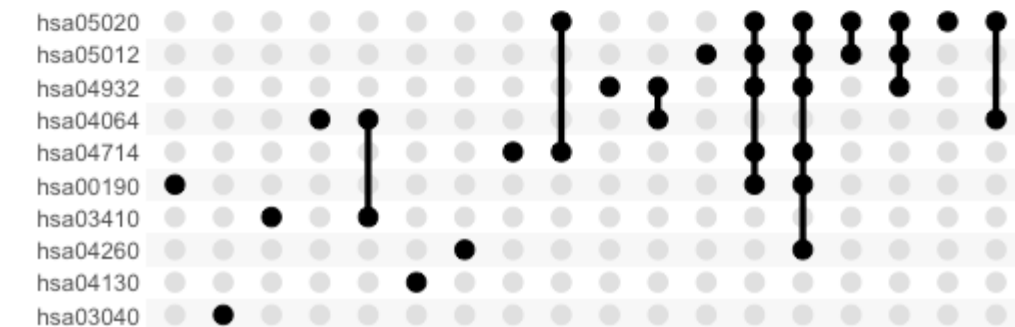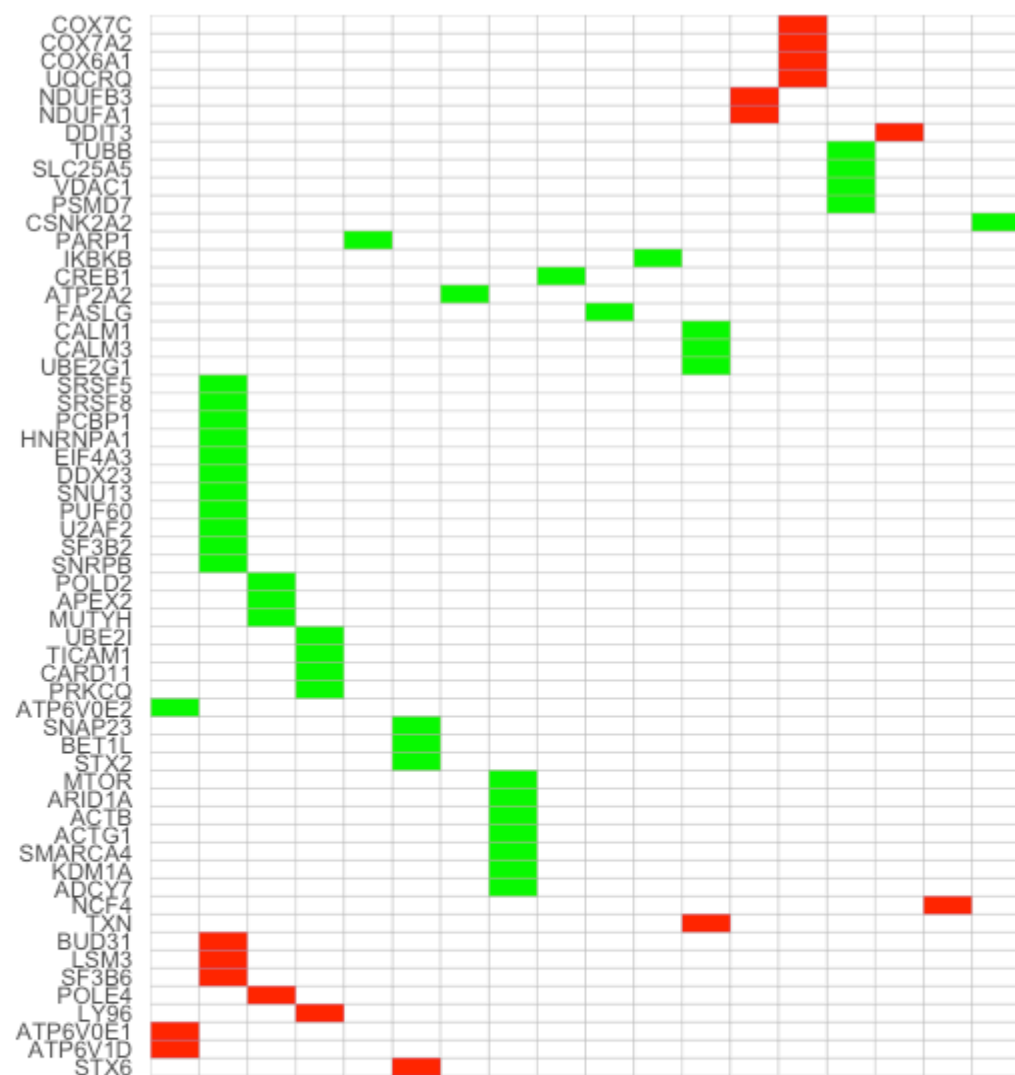*term_gene_heatmap()*

UpSet Plot

- UpSet plots are plots of the intersections of sets as a matrix
- This function creates a ggplot object of an UpSet plot where the x-axis is the UpSet plot of intersections of enriched terms

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:
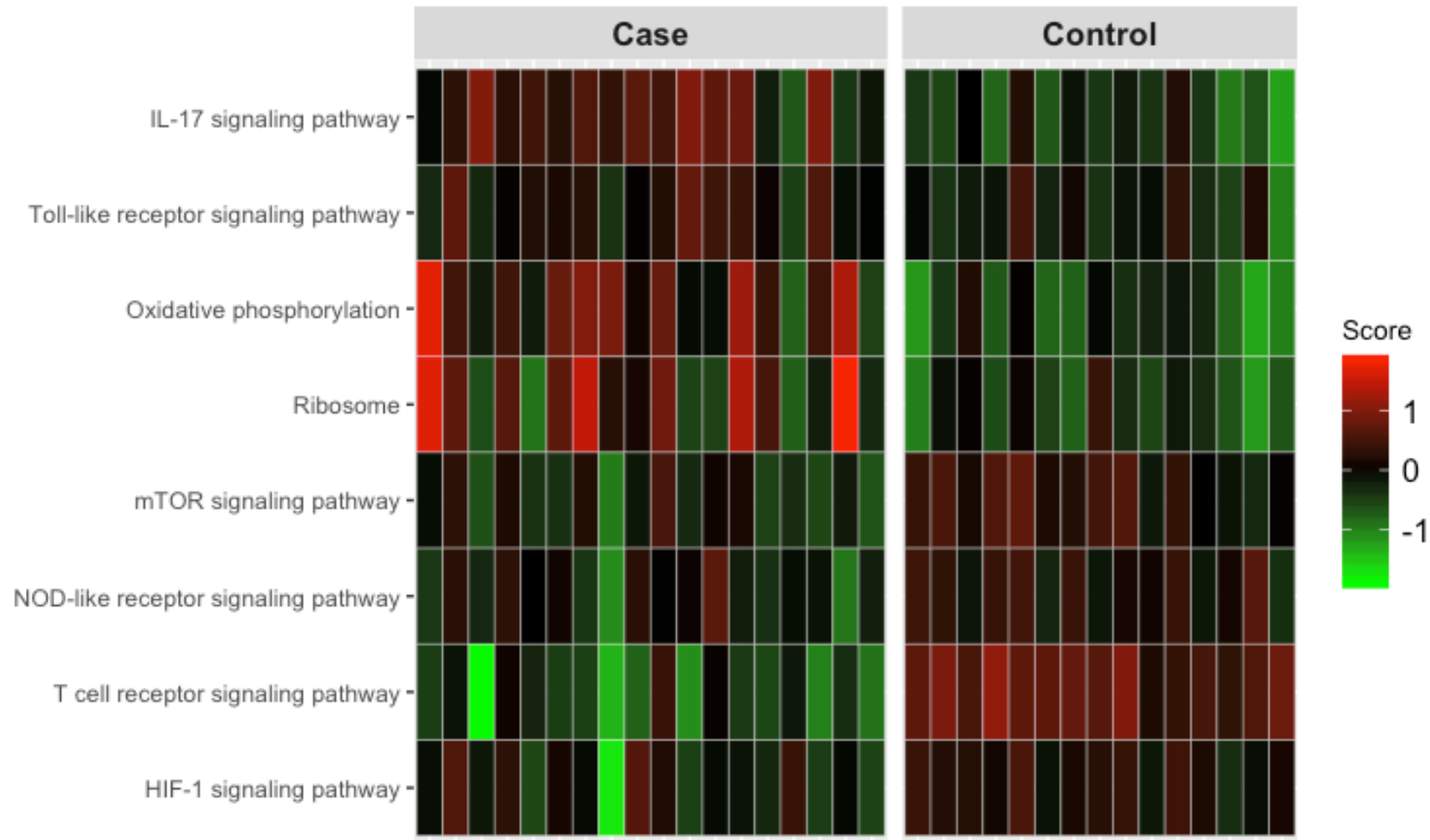
$$GS_{gs} = \frac{E_{gs} - \bar{e}_g}{s_g}$$

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:

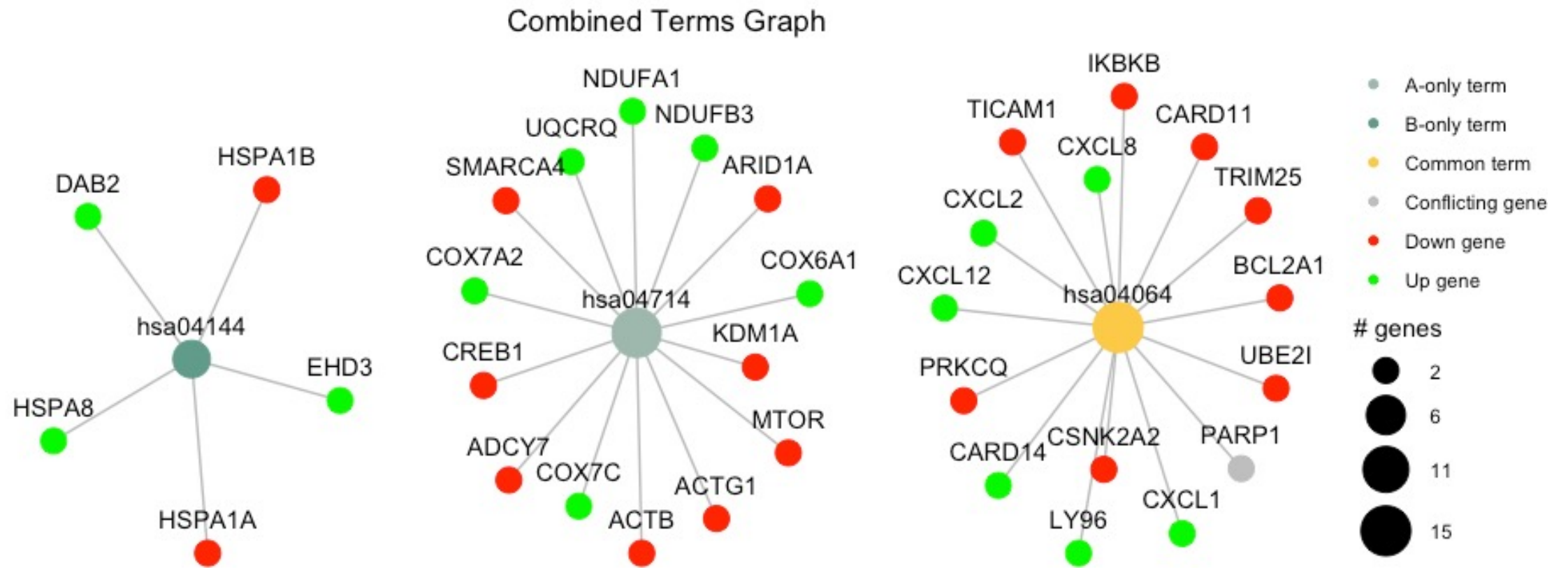$$TS_{ts} = \frac{1}{|t|} \sum_{g \in t} GS_{gs}, \text{ where } t \in T \text{ and } s \in S.$$

Heatmap of Agglomerated Scores
grouped by Case/Control

Comparing 2 pathfindR Results

combine_pathfindR_results()   *combined_results_graph()*

Combined Terms Graph

Installation (CRAN release version)

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
docker pull egeulgen/pathfindr:2.0.0
```

pathfindR

Installation (Development version)

From GitHub

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

pathfindR

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:	2.4.0
Depends:	R (≥ 4.0), pathfindR.data (≥ 2.0)
Imports:	DBI, AnnotationDbi, doParallel, foreach, rmarkdown, org.Hs.eg.db, ggplot2, ggraph, ggupset, fpc, ggkegg, grDevices, httr, igraph, R.utils, msigdbr, knitr
Suggests:	testthat (≥ 2.3.2), covr, mockery
Published:	2024-04-28
Author:	Ege Ulgen ⬤ [cre, cph], Ozan Ozisik ⬤ [aut]
Maintainer:	Ege Ulgen <egeulgen at gmail.com>
BugReports:	https://github.com/egeulgen/pathfindR/issues
License:	MIT + file LICENSE
URL:	https://egeulgen.github.io/pathfindR/, https://github.com/egeulgen/pathfindR
NeedsCompilation:	no
SystemRequirements:	Java (>= 8.0)
Citation:	pathfindR citation info
Materials:	NEWS
In views:	Omics
CRAN checks:	pathfindR results

Documentation:

Reference manual:	pathfindR.pdf
Vignettes:	Comparing Two pathfindR Results
	Introduction to pathfindR
	Step-by-Step Execution of the pathfindR Enrichment Workflow
	pathfindR Analysis for non-Homo-sapiens organisms
	Obtaining PIN and Gene Sets Data
	Visualization of pathfindR Enrichment Results

CRAN release 2.4.0

Resources

Articles   https://egeulgen.github.io/pathfindR/articles/

pathfindR website: https://egeulgen.github.io/pathfindR

To report any issues: https://github.com/egeulgen/pathfindR/issues

For all other questions: egeulgen [at] gmail.com