

# Package ‘pathfindR.data’

November 20, 2020

**Title** Data Package for 'pathfindR'

**Version** 1.1.1

**Maintainer** Ege Ulgen <egeulgen@gmail.com>

**Description** This is a data-only package, containing data needed to run the CRAN package 'pathfindR', a package for enrichment analysis utilizing active subnetworks. This package contains protein-protein interaction network data, data related to gene sets and example input/output data.

**License** MIT + file LICENSE

**Encoding** UTF-8

**LazyData** true

**Depends** R (>= 4.0)

**RoxygenNote** 7.1.1

**URL** <https://github.com/egeulgen/pathfindR.data>

**BugReports** <https://github.com/egeulgen/pathfindR.data/issues>

**NeedsCompilation** no

**Author** Ege Ulgen [cre, cph] (<<https://orcid.org/0000-0003-2090-3621>>),  
Ozan Ozisik [aut] (<<https://orcid.org/0000-0001-5980-8002>>)

**Repository** CRAN

**Date/Publication** 2020-11-20 13:50:02 UTC

## R topics documented:

biocarta_descriptions . . . . .	2
biocarta_genes . . . . .	3
biogrid_adj_list . . . . .	3
cell_markers_descriptions . . . . .	4
cell_markers_gsets . . . . .	4
custom_result . . . . .	5
example_active_snws . . . . .	5
genemania_adj_list . . . . .	6

go_all_genes . . . . .	6
GO_all_terms_df . . . . .	7
intact_adj_list . . . . .	7
kegg_adj_list . . . . .	8
kegg_descriptions . . . . .	8
kegg_genes . . . . .	9
mmu_kegg_descriptions . . . . .	9
mmu_kegg_genes . . . . .	10
mmu_string_adj_list . . . . .	10
myeloma_input . . . . .	11
myeloma_output . . . . .	11
pathfindR.data_updates . . . . .	12
RA_clustered . . . . .	13
RA_comparison_output . . . . .	14
RA_exp_mat . . . . .	15
RA_input . . . . .	15
RA_output . . . . .	16
reactome_descriptions . . . . .	17
reactome_genes . . . . .	17
string_adj_list . . . . .	18
<b>Index</b>	<b>19</b>

---

biocarta\_descriptions *BioCarta Pathways - Descriptions*

---

## Description

A named vector containing the descriptions for each human BioCarta pathway. *Generated on Sep 20, 2020.*

## Usage

```
biocarta_descriptions
```

## Format

named vector containing 289 character values, the descriptions for the given pathways.

---

biocarta_genes	<i>BioCarta Pathways - Gene Sets</i>
----------------	--------------------------------------

---

**Description**

A list containing the genes involved in each human BioCarta pathway. Each element is a vector of gene symbols located in the given pathway. *Generated on Sep 20, 2020.*

**Usage**

biocarta\_genes

**Format**

list containing 289 vectors of gene symbols. Each vector corresponds to a gene set.

---

biogrid_adj_list	<i>BioGRID PIN Adjacency List</i>
------------------	-----------------------------------

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the BioGRID protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). There are 463119 interactions in the BioGRID PIN. *Generated on Nov 6, 2020.*

**Usage**

biogrid\_adj\_list

**Format**

list containing 20192 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

cell\_markers\_descriptions

*Human Cell Markers - Descriptions*

---

### **Description**

A named vector containing descriptions of different cell types from different tissues in human. Names of the vectors are Cell Ontology IDs (if available) of the cell types in the following format: "tissue type, cancer type, cell name" For more information, refer to the article: Zhang X, Lan Y, Xu J, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 2019;47(D1):D721-D728. *Generated on Jan 16, 2020.*

### **Usage**

cell\_markers\_descriptions

### **Format**

named vector containing 496 character values, the descriptions for the given human cell types.

---

cell\_markers\_gsets

*Human Cell Markers - Gene Sets*

---

### **Description**

A list containing the sets of genes that are cell markers of different cell types from different tissues in human. Each element is a vector of cell marker gene symbols for the given cell type. Names correspond to the Cell Ontology ID (if available) of the cell type. For more information, refer to the article: Zhang X, Lan Y, Xu J, et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 2019;47(D1):D721-D728. *Generated on Jan 16, 2020.*

### **Usage**

cell\_markers\_gsets

### **Format**

list containing 496 vectors. Each vector corresponds to a cell marker gene set for a given human cell type.

---

custom_result	<i>Custom Gene Set Enrichment Results</i>
---------------	---

---

**Description**

A data frame consisting of pathfindR enrichment analysis results on the example TF target genes data (target gene sets of CREB and MYC). *Generated on Sep 20, 2020.*

**Usage**

```
custom_result
```

**Format**

data frame containing 2 rows and 8 columns. Each row is a gene set (the TF target gene sets).

---

example_active_snws	<i>Example Active Subnetworks</i>
---------------------	-----------------------------------

---

**Description**

A list of vectors containing genes for each active subnetwork that passed the filtering step. *Generated on Nov 1, 2019.*

**Usage**

```
example_active_snws
```

**Format**

list containing 112 vectors. Each vector is the set of genes for the given active subnetwork.

---

genemania\_adj\_list      *GeneMania PIN Adjacency List*

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the GeneMania protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). *Generated on Oct 12, 2019.*

**Usage**

genemania\_adj\_list

**Format**

list containing 12345 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

---

go\_all\_genes      *Gene Ontology - All Gene Ontology Gene Sets*

---

**Description**

A list containing the genes involved in each GO ontology term. Each element is a vector of gene symbols located in the given gene set. *Generated on Sep 20, 2020.*

**Usage**

go\_all\_genes

**Format**

list containing 14906 vectors of gene symbols. Each vector corresponds to a gene set.

---

GO_all_terms_df	<i>Gene Ontology - All Gene Ontology Descriptions</i>
-----------------	---

---

**Description**

A data frame containing descriptions of Gene Ontology terms (for all categories) *Generated on Sep 20, 2020.*

**Usage**

GO\_all\_terms\_df

**Format**

data frame containing 14906 rows and 3 columns. Columns are

**GO\_ID** ID of the GO term

**GO\_term** Description the GO term

**Category** Category of the GO term (i.e., "Component", "Function" or "Process")

---

intact_adj_list	<i>IntAct PIN Adjacency List</i>
-----------------	----------------------------------

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the IntAct protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). There are 262093 interactions in the IntAct PIN. *Generated on Sep 20, 2020.*

**Usage**

intact\_adj\_list

**Format**

list containing 15079 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

---

kegg_adj_list	<i>KEGG PIN Adjacency List</i>
---------------	--------------------------------

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the KEGG protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). There are 55519 interactions in the KEGG PIN. *Generated on Sep 20, 2020.*

**Usage**

```
kegg_adj_list
```

**Format**

list containing 4736 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

---

kegg_descriptions	<i>KEGG Pathways - Descriptions</i>
-------------------	-------------------------------------

---

**Description**

A named vector containing the descriptions for each Homo sapiens KEGG pathway. Names of the vector correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. *Generated on Sep 20, 2020.*

**Usage**

```
kegg_descriptions
```

**Format**

named vector containing 328 character values, the descriptions for the given pathways.

---

`kegg_genes`*KEGG Pathways - Gene Sets*

---

**Description**

A list containing the genes involved in each Homo sapiens KEGG pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. *Generated on Sep 20, 2020.*

**Usage**`kegg_genes`**Format**

list containing 328 vectors of gene symbols. Each vector corresponds to a pathway.

---

`mmu_kegg_descriptions` *Mus Musculus KEGG Pathways - Descriptions*

---

**Description**

A named vector containing the descriptions for each Mus musculus KEGG pathway. Names of the vector correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. *Generated on Sep 20, 2020.*

**Usage**`mmu_kegg_descriptions`**Format**

named vector containing 324 character values, the descriptions for the given pathways.

---

mmu\_kegg\_genes      *Mus Musculus KEGG Pathways - Gene Sets*

---

**Description**

A list containing the genes involved in each Mus musculus KEGG pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the KEGG ID of the pathway. Pathways that did not contain any genes were discarded. *Generated on Sep 20, 2020.*

**Usage**

mmu\_kegg\_genes

**Format**

list containing 324 vectors of gene symbols. Each vector corresponds to a pathway.

---

mmu\_string\_adj\_list      *Mus musculus STRING PIN Adjacency List*

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the Mus musculus STRING protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). Only interactions with a combined score  $\geq 800$  were kept. *Generated on Nov 1, 2019.*

**Usage**

mmu\_string\_adj\_list

**Format**

list containing 11217 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

---

`myeloma_input`*Example Input for Myeloma Analysis (Mus Musculus)*

---

**Description**

A dataset containing the differentially-expressed genes and adjusted p-values for the GEO dataset GSE99393. The RNA microarray experiment was performed to detail the global program of gene expression underlying polarization of myeloma-associated macrophages by CSF1R antibody treatment. The samples were 6 murine bone marrow derived macrophages co-cultured with myeloma cells (myeloma-associated macrophages), 3 of which were treated with CSF1R antibody (treatment group) and the rest were treated with control IgG antibody (control group). In this dataset, differentially-expressed genes with  $|\log_{2}FC| \geq 2$  and  $FDR < 0.05$  are presented. *Generated on Nov 1, 2019.*

**Usage**`myeloma_input`**Format**

A data frame with 45 rows and 2 variables:

**Gene\_Symbol** MGI gene symbols of the differentially-expressed genes

**FDR** adjusted p values, via the Benjamini & Hochberg (1995) method

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE99393>

**See Also**

[myeloma\\_output](#) for the example mmu enrichment output. [run\\_pathfindR](#) for details on the pathfindR enrichment analysis.

---

`myeloma_output`*Example Output for Myeloma Analysis (Mus Musculus)*

---

**Description**

A dataset containing the results of pathfindR's active-subnetwork-oriented enrichment workflow performed on the Mus musculus myeloma differential expression dataset [myeloma\\_input](#). *Generated on Oct 19, 2020.*

**Usage**`myeloma_output`

**Format**

A data frame with 18 rows and 9 columns:

**ID** ID of the enriched term

**Term\_Description** Description of the enriched term

**Fold\_Enrichment** Fold enrichment value for the enriched term

**occurrence** the number of iterations that the given term was found to enriched over all iterations

**support** the median support (proportion of active subnetworks leading to enrichment within an iteration) over all iterations

**lowest\_p** the lowest adjusted-p value of the given term over all iterations

**highest\_p** the highest adjusted-p value of the given term over all iterations

**Up\_regulated** the up-regulated genes in the input involved in the given term, comma-separated

**Down\_regulated** the down-regulated genes in the input involved in the given term, comma-separated

**See Also**

[myeloma\\_input](#) for the example mmu input. [run\\_pathfindR](#) for details on the pathfindR enrichment workflow.

---

pathfindR.data\_updates

*Table of Data for pathfindR*

---

**Description**

Data frame containing all the data for pathfindR along with descriptions and last update dates.

**Usage**

```
pathfindR.data_updates
```

**Format**

A data frame with 30 rows and 4 columns:

**Category** Category of the data

**Name** Name of the data

**Description** Description of the data

**Last\_Update** Last update date

---

RA_clustered	<i>Example Output for the pathfindR Clustering Workflow - Rheumatoid Arthritis</i>
--------------	--

---

## Description

A dataset containing the results of pathfindR's clustering and partitioning workflow performed on the rheumatoid arthritis enrichment results [RA\\_output](#). The clustering and partitioning function [cluster\\_enriched\\_terms](#) was used with the default settings (i.e. hierarchical clustering was performed and the agglomeration method was "average"). *Generated on Oct 19, 2020.*

## Usage

```
RA_clustered
```

## Format

A data frame with 95 rows and 11 columns:

**ID** ID of the enriched term

**Term\_Description** Description of the enriched term

**Fold\_Enrichment** Fold enrichment value for the enriched term

**occurrence** the number of iterations that the given term was found to enriched over all iterations

**support** the median support (proportion of active subnetworks leading to enrichment within an iteration) over all iterations

**lowest\_p** the lowest adjusted-p value of the given term over all iterations

**highest\_p** the highest adjusted-p value of the given term over all iterations

**Up\_regulated** the up-regulated genes in the input involved in the given term, comma-separated

**Down\_regulated** the down-regulated genes in the input involved in the given term, comma-separated

**Cluster** the cluster to which the enriched term is assigned

**Status** whether the enriched term is the "Representative" term in its cluster or only a "Member"

## See Also

[RA\\_input](#) for the RA differentially-expressed genes data frame [RA\\_clustered](#) for the RA example pathfindR clustering output [RA\\_exp\\_mat](#) for the RA differentially-expressed genes expression matrix [run\\_pathfindR](#) for details on the pathfindR enrichment analysis [RA\\_output](#) for the RA example pathfindR enrichment output [cluster\\_enriched\\_terms](#) for details on clustering methods

---

RA\_comparison\_output *Second Example Output for the pathfindR Enrichment Workflow*

---

### Description

The data frame containing the results of pathfindR's active-subnetwork-oriented enrichment workflow performed on the rheumatoid arthritis dataset GSE84074 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84074>. Analysis via run\_pathfindR was performed using the default settings. *Generated on Oct 19, 2020.*

### Usage

```
RA_comparison_output
```

### Format

A data frame with 35 rows and 9 columns:

**ID** ID of the enriched term

**Term\_Description** Description of the enriched term

**Fold\_Enrichment** Fold enrichment value for the enriched term

**occurrence** the number of iterations that the given term was found to enriched over all iterations

**support** the median support (proportion of active subnetworks leading to enrichment within an iteration) over all iterations

**lowest\_p** the lowest adjusted-p value of the given term over all iterations

**highest\_p** the highest adjusted-p value of the given term over all iterations

**Up\_regulated** the up-regulated genes in the input involved in the given term, comma-separated

**Down\_regulated** the down-regulated genes in the input involved in the given term, comma-separated

### See Also

[RA\\_input](#) for the RA differentially-expressed genes data frame [RA\\_output](#) for the RA example pathfindR enrichment output [RA\\_clustered](#) for the RA example pathfindR clustering output [RA\\_exp\\_mat](#) for the RA differentially-expressed genes expression matrix [run\\_pathfindR](#) for details on the pathfindR enrichment analysis

---

`RA_exp_mat`*Example Input for pathfindR - Enriched Term Scoring*

---

**Description**

A matrix containing the  $\log_2$ -transformed and quantile-normalized expression values of the differentially-expressed genes for 18 rheumatoid arthritis (RA) patients and 15 healthy subjects. The matrix contains expression values of 572 significantly differentially-expressed genes (see [RA\\_input](#)) with  $\text{adj.P.Val} \leq 0.05$ . *Generated on Sep 28, 2019.*

**Usage**`RA_exp_mat`**Format**

A matrix with 572 rows and 33 columns.

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15573>

**See Also**

[RA\\_input](#) for the RA differentially-expressed genes data frame [RA\\_output](#) for the RA example pathfindR enrichment output [score\\_terms](#) for details on calculating agglomerated scores of enriched terms

---

`RA_input`*Example Input for the pathfindR Enrichment Workflow - Rheumatoid Arthritis*

---

**Description**

A dataset containing the differentially-expressed genes along with the associated  $\log_2$ (fold-change) values and FDR adjusted p-values for the GEO dataset GSE15573. This microarray dataset aimed to characterize gene expression profiles in the peripheral blood mononuclear cells of 18 rheumatoid arthritis (RA) patients versus 15 healthy subjects. Differentially-expressed genes with  $\text{adj.P.Val} < 0.05$  are presented in this data frame. *Generated on Nov 1, 2019.*

**Usage**`RA_input`

**Format**

A data frame with 572 rows and 3 variables:

**Gene.symbol** HGNC gene symbols of the differentially-expressed genes

**logFC**  $\log_2$ (fold-change) values

**adj.P.Val** adjusted p values, via the Benjamini & Hochberg (1995) method

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15573>

**See Also**

[RA\\_output](#) for the RA example pathfindR enrichment output [RA\\_clustered](#) for the RA example pathfindR clustering output [RA\\_exp\\_mat](#) for the RA differentially-expressed genes expression matrix [run\\_pathfindR](#) for details on the pathfindR enrichment analysis

---

RA\_output

*Example Output for the pathfindR Enrichment Workflow - Rheumatoid Arthritis*

---

**Description**

The data frame containing the results of pathfindR's active-subnetwork-oriented enrichment workflow performed on the rheumatoid arthritis differential-expression data frame [RA\\_input](#). Analysis via `run_pathfindR` was performed using the default settings. *Generated on Oct 19, 2020.*

**Usage**

RA\_output

**Format**

A data frame with 95 rows and 9 columns:

**ID** ID of the enriched term

**Term\_Description** Description of the enriched term

**Fold\_Enrichment** Fold enrichment value for the enriched term

**occurrence** the number of iterations that the given term was found to enriched over all iterations

**support** the median support (proportion of active subnetworks leading to enrichment within an iteration) over all iterations

**lowest\_p** the lowest adjusted-p value of the given term over all iterations

**highest\_p** the highest adjusted-p value of the given term over all iterations

**Up\_regulated** the up-regulated genes in the input involved in the given term, comma-separated

**Down\_regulated** the down-regulated genes in the input involved in the given term, comma-separated

**See Also**

[RA\\_input](#) for the RA differentially-expressed genes data frame [RA\\_clustered](#) for the RA example pathfindR clustering outputs [RA\\_exp\\_mat](#) for the RA differentially-expressed genes expression matrix [run\\_pathfindR](#) for details on the pathfindR enrichment analysis

---

reactome\_descriptions *Reactome Pathways - Descriptions*

---

**Description**

A named vector containing the descriptions for each human Reactome pathway. Names of the vector correspond to the Reactome ID of the pathway. *Generated on Sep 20, 2020.*

**Usage**

```
reactome_descriptions
```

**Format**

named vector containing 2400 character values, the descriptions for the given pathways.

---

reactome\_genes *Reactome Pathways - Gene Sets*

---

**Description**

A list containing the genes involved in each human Reactome pathway. Each element is a vector of gene symbols located in the given pathway. Names correspond to the Reactome ID of the pathway. *Generated on Sep 20, 2020.*

**Usage**

```
reactome_genes
```

**Format**

list containing 2400 vectors of gene symbols. Each vector corresponds to a pathway.

---

string_adj_list	<i>STRING PIN Adjacency List</i>
-----------------	----------------------------------

---

**Description**

An adjacency list of vectors containing interactors B for each interactor A in the STRING protein-protein interaction network (The designations "interactor A" and "interactor B" are arbitrary). Only interactions with a combined score  $\geq 800$  were kept. *Generated on Oct 31, 2019.*

**Usage**

```
string_adj_list
```

**Format**

list containing 11934 vectors. Each vector is the set of gene symbols of interactors B for each interactor A.

# Index

## \* datasets

- biocarta\_descriptions, [2](#)
  - biocarta\_genes, [3](#)
  - biogrid\_adj\_list, [3](#)
  - cell\_markers\_descriptions, [4](#)
  - cell\_markers\_gsets, [4](#)
  - custom\_result, [5](#)
  - example\_active\_snws, [5](#)
  - genemania\_adj\_list, [6](#)
  - go\_all\_genes, [6](#)
  - GO\_all\_terms\_df, [7](#)
  - intact\_adj\_list, [7](#)
  - kegg\_adj\_list, [8](#)
  - kegg\_descriptions, [8](#)
  - kegg\_genes, [9](#)
  - mmu\_kegg\_descriptions, [9](#)
  - mmu\_kegg\_genes, [10](#)
  - mmu\_string\_adj\_list, [10](#)
  - myeloma\_input, [11](#), [11](#), [12](#)
  - myeloma\_output, [11](#), [11](#)
  - pathfindR.data\_updates, [12](#)
  - RA\_clustered, [13](#), [13](#), [14](#), [16](#), [17](#)
  - RA\_comparison\_output, [14](#)
  - RA\_exp\_mat, [13](#), [14](#), [15](#), [16](#), [17](#)
  - RA\_input, [13–15](#), [15](#), [16](#), [17](#)
  - RA\_output, [13–16](#), [16](#)
  - reactome\_descriptions, [17](#)
  - reactome\_genes, [17](#)
  - run\_pathfindR, [11–14](#), [16](#), [17](#)
  - score\_terms, [15](#)
  - string\_adj\_list, [18](#)
- 
- biocarta\_descriptions, [2](#)
  - biocarta\_genes, [3](#)
  - biogrid\_adj\_list, [3](#)
  - cell\_markers\_descriptions, [4](#)
  - cell\_markers\_gsets, [4](#)
  - cluster\_enriched\_terms, [13](#)
  - custom\_result, [5](#)
  - example\_active\_snws, [5](#)
  - genemania\_adj\_list, [6](#)
  - go\_all\_genes, [6](#)
  - GO\_all\_terms\_df, [7](#)
  - intact\_adj\_list, [7](#)
  - kegg\_adj\_list, [8](#)
  - kegg\_descriptions, [8](#)
  - kegg\_genes, [9](#)
  - mmu\_kegg\_descriptions, [9](#)
  - mmu\_kegg\_genes, [10](#)
  - mmu\_string\_adj\_list, [10](#)
  - myeloma\_input, [11](#), [11](#), [12](#)
  - myeloma\_output, [11](#), [11](#)
  - pathfindR.data\_updates, [12](#)
  - RA\_clustered, [13](#), [13](#), [14](#), [16](#), [17](#)
  - RA\_comparison\_output, [14](#)
  - RA\_exp\_mat, [13](#), [14](#), [15](#), [16](#), [17](#)
  - RA\_input, [13–15](#), [15](#), [16](#), [17](#)
  - RA\_output, [13–16](#), [16](#)
  - reactome\_descriptions, [17](#)
  - reactome\_genes, [17](#)
  - run\_pathfindR, [11–14](#), [16](#), [17](#)
  - score\_terms, [15](#)
  - string\_adj\_list, [18](#)