

# Package ‘SEMdata’

June 22, 2022

**Type** Package

**Title** Datasets for SEM-based network analysis using the SEMgraph R package

**Version** 1.1.0

**Date** 2022-06-22

**Author** Mario Grassi [aut], Fernando Palluzzi [aut, cre]

**Maintainer** Fernando Palluzzi <fernando.palluzzi@gmail.com>

**Description** This package contains a collection of datasets and reference interaction networks for graph analysis in computational biology.

**License** GPL-3

**Encoding** UTF-8

**LazyData** TRUE

**Depends** R (>= 4.0)

**Suggests** BiocStyle, graphite, igraph, knitr, org.Hs.eg.db, rmarkdown, SEMgraph

**VignetteBuilder** knitr

**RoxygenNote** 7.1.2

**NeedsCompilation** no

## R topics documented:

alsData . . . . .	2
covid19 . . . . .	3
ftdDName . . . . .	3
reactome . . . . .	4
reactome.pathways . . . . .	5
string . . . . .	6
<b>Index</b>	<b>8</b>

---

alsData

*Amyotrophic Lateral Sclerosis (ALS) dataset*


---

## Description

Expression profiling through high-throughput sequencing (RNA-seq) of 139 ALS patients and 21 healthy controls (HCs), from Tam et al. (2019).

## Usage

```
alsData
```

## Format

alsData is a list of 4 objects:

1. "graph", ALS graph as the largest connected component of the "Amyotrophic lateral sclerosis (ALS)" pathway from KEGG database;
2. "exprs", a matrix of 160 rows (subjects) and 17695 columns (genes). Raw data from the GEO dataset GSE124439 (Tam et al., 2019) were pre-processed applying batch effect correction, using the sva R package (Leek et al., 2012), to remove data production center and brain area biases. Using multidimensional scaling-based clustering, ALS-specific and an HC-specific clusters were generated. Misclassified samples were blacklisted and removed from the current dataset;
3. "group", a binary group vector of 139 ALS subjects (1) and 21 healthy controls (0);
4. "details", a data.frame reporting information about included and blacklisted samples.

## Source

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124439>

## References

Tam OH, Rozhkov NV, Shaw R, Kim D et al. (2019). Postmortem Cortex Samples Identify Distinct Molecular Subtypes of ALS: Retrotransposon Activation, Oxidative Stress, and Activated Glia. *Cell Reppts*, 29(5):1164-1177.e5. <<https://doi.org/10.1016/j.celrep.2019.09.066>>

Jeffrey T. Leek, W. Evan Johnson, Hilary S. Parker, Andrew E. Jaffe, and John D. Storey (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. Mar 15; 28(6): 882-883. <<https://doi.org/10.1093/bioinformatics/bts034>>

## Examples

```
alsData$graph
dim(alsData$exprs)
table(alsData$group)
```

---

covid19*COVID-19 RNA-seq dataset*

---

**Description**

Expression profiling through high-throughput sequencing (RNA-seq, Illumina NovaSeq 6000) of 72 young human patients (< 50 years old) with severe forms of COVID-19, from Carapito et al. (2022).

**Usage**

```
covid19
```

**Format**

covis19 is a data.frame of 15957 rows (genes) and 69 columns (subjects), including RSEM count data for 23 (C1-23) non-critical and 46 (P1-46) critical COVID-19 cases. The former group is composed by patients in a non-critical care ward under supplemental oxygen, while the latter included patients in the intensive care unit with Acute Respiratory Distress Syndrome (ARDS), as reported by Carapito et al. 2022.

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE172114>

**References**

Carapito R, Li R, Helms J, Carapito C et al. Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort. Sci Transl Med 2022 Jan 19;14(628):eabj7521. <<https://doi.org/10.1126/scitranslmed.abj7521>>

**Examples**

```
summary(covid19)
dim(covid19)
```

---

ftdDNAm*Frontotemporal Dementia (FTD) DNA methylation dataset*

---

**Description**

DNA methylation (DNAm) profiling by genome tiling array of 105 FTD patients and 150 healthy controls from peripheral blood samples (Li et al., 2014).

**Usage**

```
ftdDName
```

**Format**

ftdDName is a list of 2 objects:

1. "pc1", a data matrix of 256 rows (subjects) and 16560 columns (genes) containing the value of the first principal component of DName levels, obtained applying a principal component analysis to methylated CpG sites within the promoter region, for each gene (genes with unmethylated CpGs in both conditions were discarded);
2. "group", a binary group vector of 105 FTD patients (1) and 150 healthy controls (0).

**Source**

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53740>

**References**

Li Y, Chen JA, Sears RL, Gao F et al. An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. PLoS Genet 2014 Mar;10(3):e1004211. <<https://doi.org/10.1371/journal.pgen.1004211>>

**Examples**

```
dim(ftdDName$pc1)
table(ftdDName$group)
```

---

reactome

*Reactome interactome*

---

**Description**

Interactome generated by merging Reactome pathways extracted using the graphite R package (update: April, 2020).

**Usage**

```
reactome
```

**Format**

"reactome" is an igraph network object of 9762 nodes and 416128 edges corresponding to the union of 1641 Reactome pathways.

**Source**

<https://reactome.org>

## References

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498-D503. PubMed PMID: 31691815. <doi:10.1093/nar/gkz1031>

Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 41(1):e19. <https://doi.org/10.1093/nar/gks866>

## Examples

```
summary(reactome)

# Reactome degrees of freedom
vcount(reactome)*(vcount(reactome) - 1)/2 - ecount(reactome)

# Reactome average shortest path length
mean_distance(reactome)
```

---

reactome.pathways	<i>Reactome pathways collection</i>
-------------------	-------------------------------------

---

## Description

Collection of Reactome signaling pathways extracted using the graphi te R package (update: April, 2020).

## Usage

```
reactome.pathways
```

## Format

"reactome.pathways" is a list of 1641 Reactome pathways, stored as igraph objects.

## Source

<https://reactome.org>

## References

Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucleic Acids Res. 2020 Jan 8;48(D1):D498-D503. <doi:10.1093/nar/gkz1031. PubMed PMID: 31691815.>

Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 41(1):e19. <https://doi.org/10.1093/nar/gks866>

## Examples

```
## Not run:

# Number of nodes per pathway
react.nodes <- unlist(lapply(reactome.pathways, vcount))
names(react.nodes) <- NULL
# Number of edges per pathway
react.edges <- unlist(lapply(reactome.pathways, ecount))
names(react.edges) <- NULL
# Gene list per pathway
react.genes <- unlist(lapply(reactome.pathways, function(x) V(x)$name))
quantile(react.nodes)
quantile(react.edges)
length(unique(react.genes)) # Number of unique genes within the dataset

# Loading breast cancer KEGG network
i <- which(names(reactome.pathways) == "NOTCH4 Intracellular Domain Regulates Transcription")
notch4.graph <- reactome.pathways[[i]]
summary(notch4.graph)
gplot(notch4.graph)

## End(Not run)
```

---

string

*STRING interactome*

---

## Description

STRING interactome version 10.5.

## Usage

string

## Format

"string" is an igraph network object of 9725 nodes and 170987 edges corresponding to the STRING interactome (version 10.5).

## Source

<https://string-db.org>

**References**

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, von Mering C (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47: D607-613. <<https://doi.org/10.1093/nar/gky1131>>

**Examples**

```
string
summary(string)

# STRING degrees of freedom
vcount(string)*(vcount(string) - 1)/2 - ecound(string)

# STRING average shortest path length
mean_distance(string)
```

# Index

alsData, [2](#)

covid19, [3](#)

ftdDName, [3](#)

reactome, [4](#)

reactome.pathways, [5](#)

string, [6](#)