Package: clustermole (via r-universe)

October 11, 2026

Type Package

Title Cell Type Marker Database for Single-Cell RNA-Seq Data

Version 1.2.0

Description Provides a meta-database of thousands of human and mouse cell identity markers curated from multiple sources, along with methods for cell type prediction based on marker gene overlaps or gene set enrichment.

License MIT + file LICENSE

URL https://igordot.github.io/clustermole/

BugReports https://github.com/igordot/clustermole/issues

Depends R (>= 4.3)

Imports dplyr (>= 1.1.0), methods, rlang, stats, tibble, tidyr, utils

Suggests covr, GSEABase, GSVA (>= 1.50.0), knitr, rmarkdown, roxygen2, singscore, testthat

Config/roxygen2/version 8.1.0

Encoding UTF-8

Roxygen list(markdown = TRUE)

Config/pak/sysreqs libicu-dev

Repository https://igordot.r-universe.dev

Date/Publication 2026-10-07 02:23:37 UTC

RemoteUrl https://github.com/igordot/clustermole

RemoteRef v1.2.0

RemoteSha 2eeeea5d6335070b395b8b8d57f7d4867acb3f82

Contents

clustermole_enrichment 2
clustermole_markers 3
clustermole_overlaps 4
read_gmt 5

Index 6

clustermole_enrichment

Cell types based on the expression of all genes

Description

Score cell type signatures using the full gene expression matrix.

Usage

```
clustermole_enrichment(expr_mat, species, method = "gsva", max_rank = 100)
```

Arguments

`expr_mat`	Numeric matrix or data frame of logCPMs or logTPMs. Must contain at least 5,000 gene rows and five cluster/population columns. Row names must be unique.
`species`	Gene symbol species: `hs` for human or `mm` for mouse.
`method`	Enrichment method: `gsva` (default), `ssgsea`, `singscore`, or `all` to combine ranks from all three methods. See references below.
`max_rank`	Maximum signature rank to return. Defaults to `100`.

Value

A data frame with one row per returned signature and input column:

- `cluster`: Input column name.
- `score`: Enrichment score (higher means greater enrichment).
- `score_rank`: Signature rank (lower means greater enrichment). With `method = "all"`, this is the average rank across methods.
- Signature metadata (see `clustermole_markers()`).

With `method = "all"`, these columns replace `score`:

- `score_rank_{method}`: The ranks from each method.
- `score_ranks_{stat}`: Minimum, mean, and median ranks across methods.

References

Barbie, D., Tamayo, P., Boehm, J. et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108–112 (2009). doi:10.1038/nature08460

Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14, 7 (2013). doi:10.1186/14712105147

Foroutan, M., Bhuva, D.D., Lyu, R. et al. Single sample scoring of molecular phenotypes. *BMC Bioinformatics* 19, 404 (2018). doi:10.1186/s1285901824354

Examples

```
# my_enrichment <- clustermole_enrichment(
#   expr_mat = my_expr_mat, species = "hs"
# )
```

---

clustermole_markers *Available cell type markers*

Description

Retrieve cell type markers from the clustermole database.

Usage

```
clustermole_markers(species = c("hs", "mm"))
```

Arguments

species	Gene symbol species: hs for human or mm for mouse.

Details

The gene column uses the official NCBI symbols for the requested species. The package maps the aliases in each source database to these symbols. The gene_original column preserves the symbol from the source. The source databases include both human and mouse cell type markers. The package maps them across species with ortholog data from the Alliance of Genome Resources and keeps only reciprocal best matches. The output does not include genes that have no ortholog in the requested species and does not include signatures with fewer than five genes.

Value

A data frame of cell type markers with these columns:

- gene: Canonical gene symbol for the requested species.
- gene_original: Original source gene symbol.
- celltype_full: Full cell type signature identifier.
- db: Source database.
- celltype: Cell type label.
- organ: Organ label.
- species: Source signature species, if known.
- n_genes: Gene count per signature.

Examples

```
markers <- clustermole_markers()
head(markers)
```

clustermole_overlaps *Cell types based on overlap of marker genes*

Description

Perform overrepresentation analysis for a set of genes compared to all cell type signatures.

Usage

```
clustermole_overlaps(genes, species, max_p = 0.05, max_fdr = 1)
```

Arguments

genes	A character vector of gene symbols.
species	Gene symbol species: hs for human or mm for mouse.
max_p	Maximum p-value to return. Defaults to 0.05.
max_fdr	Maximum FDR to return. Defaults to 1.

Value

A data frame with one row per returned signature:

- overlap: Unique gene count shared by the input and signature.
- p_value: Hypergeometric test p-value.
- fdr: Benjamini-Hochberg adjusted p-value across all tested signatures.
- n_genes: Unique gene count in the signature for the requested species.
- Signature metadata (see clustermole_markers()).

Examples

```
my_genes <- c("CD2", "CD3D", "CD3E", "CD3G", "TRAC", "TRBC2", "LTB")
my_overlaps <- clustermole_overlaps(genes = my_genes, species = "hs")
head(my_overlaps)
```

read_gmt *Read a GMT file into a data frame*

Description

Read a GMT file into a data frame

Usage

```
read_gmt(file, geneset_label = "celltype", gene_label = "gene")
```

Arguments

file	A file path, URL, or connection.
geneset_label	Output column name for gene sets (GMT column 1).
gene_label	Output column name for genes (GMT columns 3 onward).

Value

A data frame with gene sets and genes, one gene per row.

Examples

```
## Not run:
gmt <- "http://software.broadinstitute.org/gsea/msigdb/supplemental/scsig.all.v1.0.symbols.gmt"
gmt_tbl <- read_gmt(gmt)
head(gmt_tbl)

## End(Not run)
```

Index

clustermole_enrichment, 2
clustermole_markers, 3
clustermole_markers(), 2, *4*
clustermole_overlaps, 4

read_gmt, 5