

Package: rsemflow (via r-universe)

October 22, 2026

Type Package

Title Metadata-Driven Downstream Analysis for RSEM Gene-Level Results

Version 0.1.1

Description A command-line and R interface for reproducible downstream analysis of RSEM gene-level expression estimates. 'rsemflow' imports '*.genes.results' files together with ordinary sample metadata, computes expression-level summaries, normalization, principal component analysis, sample correlation, 'DESeq2' differential expression, preranked gene set enrichment analysis, 'GSVA' or 'ssGSEA' pathway scores, and Ensembl gene annotation. Every result is written as a tab-separated table; the package does not draw figures.

License MIT + file LICENSE

URL <https://daikikumakura.github.io/rsemflow/>,
<https://github.com/DaikiKumakura/rsemflow>

BugReports <https://github.com/DaikiKumakura/rsemflow/issues>

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

biocViews RNASeq, DifferentialExpression, GeneSetEnrichment,
Transcriptomics

Depends R (>= 4.3.0)

Imports AnnotationDbi, BiocParallel, DESeq2, GSVA (>= 2.0.0),
SummarizedExperiment, data.table, fgsea, msigdb (>= 10.0.0),
stats, tools, tximport, utils

Suggests knitr, org.Hs.eg.db, org.Mm.eg.db, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Config/testthat/edition 3

Config/pak/sysreqs libmagick++-dev gsfonts libicu-dev libpng-dev libxml2-dev libssl-dev zlib1g-dev

Repository <https://daikikumakura.r-universe.dev>
Date/Publication 2026-10-18 23:48:59 UTC
RemoteUrl <https://github.com/DaikiKumakura/rsemflow>
RemoteRef HEAD
RemoteSha 66f38bb6c804293d672d63fd4d94f3b03fdf6112

Contents

annotate_ensembl	2
annotate_gtf	3
cli_main	3
compute_correlation	4
compute_pca	4
differential_deseq2	5
normalize_expression	7
read_rsem_study	7
read_study	8
run_gsea	9
run_gsva	10
summarize_study	11
Index	12

annotate_ensembl	<i>Annotate Ensembl gene IDs using a local Bioconductor OrgDb</i>
------------------	---

Description

Adds the stable Ensembl ID (version suffix removed), gene symbol, Entrez ID, and description from org.Hs.eg.db or org.Mm.eg.db. The original RSEM gene_id is always kept. When an Ensembl ID maps to several Entrez records, the record with the most filled fields is kept. annotation_status is mapped when the OrgDb returned at least one of the symbol, Entrez ID, or description, and unmapped otherwise (including non-Ensembl IDs such as spike-ins).

Usage

```
annotate_ensembl(study, species = "auto")
```

Arguments

study	A rsemflow_study or study path.
species	auto, human, or mouse.

Value

A gene annotation data frame.

annotate_gtf	<i>Annotate genes from a GTF file</i>
--------------	---------------------------------------

Description

Reads the gene records of the GTF used to build the RSEM reference and returns gene_name, gene_type/gene_biotype, and description. IDs are matched on the full versioned ID first and on the stable Ensembl ID second.

Usage

```
annotate_gtf(study, gtf)
```

Arguments

study	A rsemflow_study or study path.
gtf	Path to a GTF or GTF.GZ file.

Value

A gene annotation data frame.

cli_main	<i>RSEMflow command-line entry point</i>
----------	--

Description

Parses rsemflow <module> <action> [options] and calls the matching R function. Options use the prefixes --i- (inputs), --m- (metadata), --p- (parameters), and --o- (outputs). Add --help at any level for usage.

Usage

```
cli_main(args = commandArgs(trailingOnly = TRUE))
```

Arguments

args	Character vector of command-line arguments.
------	---

Details

The installed launcher script (system.file("exec", "rsemflow", package = "rsemflow")) calls this function. Without the launcher, use Rscript -e "rsemflow::cli_main()" <module> <action> [options].

Value

Invisibly returns zero on success. Errors are raised as R errors.

Examples

```
cli_main("--help")
cli_main(c("differential", "deseq2", "--help"))
```

compute_correlation	<i>Compute sample correlation</i>
---------------------	-----------------------------------

Description

Compute sample correlation

Usage

```
compute_correlation(
  study,
  transform = c("vst", "log2-tpm"),
  method = c("pearson", "spearman")
)
```

Arguments

study	A rsemflow_study or study path.
transform	vst or log2-tpm.
method	pearson or spearman.

Value

A sample-by-sample correlation matrix.

compute_pca	<i>Compute PCA tables from transformed expression</i>
-------------	---

Description

Selects the ntop genes with the highest variance after transformation and runs `stats::prcomp()` with samples as observations, centering on, and no scaling.

Usage

```
compute_pca(
  study,
  transform = c("vst", "log2-tpm"),
  ntop = 500L,
  components = 10L
)
```

Arguments

study	A rsemflow_study or path.
transform	vst or log2-tpm.
ntop	Number of most variable genes to use.
components	Maximum number of PCs to return.

Value

A list containing scores, loadings, and variance data frames, plus the transform and the number of genes used (ntop).

Examples

```
ex <- system.file("extdata", "example", package = "rsemflow")
study <- read_rsem_study(file.path(ex, "data"), file.path(ex, "metadata.tsv"))
p <- compute_pca(study, ntop = 100, components = 3)
p$variance
```

differential_deseq2 *Run metadata-driven DESeq2 differential expression*

Description

Fits a DESeq2 model from the tximport-derived counts of a study and returns one result table per comparison.

Usage

```
differential_deseq2(
  study,
  formula,
  term,
  metadata = NULL,
  reference = character(),
  contrast = character(),
  at = character(),
  test = c("wald", "lrt"),
  reduced = NULL,
  categorical = character(),
  numeric = character(),
  alpha = 0.05
)
```

Arguments

<code>study</code>	A <code>rsemflow_study</code> or study path.
<code>formula</code>	R formula or formula string.
<code>term</code>	Metadata variable or interaction term to report.
<code>metadata</code>	Optional replacement metadata path/data.frame.
<code>reference</code>	Character vector like <code>condition::Control</code> .
<code>contrast</code>	Character vector like <code>condition::Drug::Control</code> .
<code>at</code>	Optional conditioning specification like <code>genotype::KO</code> ; currently supports one simple-effect condition with an explicit contrast.
<code>test</code>	<code>wald</code> or <code>lrt</code> .
<code>reduced</code>	Reduced formula required for LRT.
<code>categorical</code>	Metadata columns to force categorical.
<code>numeric</code>	Metadata columns to force numeric.
<code>alpha</code>	Significance threshold passed to DESeq2 results.

Details

Which tables are produced depends on `term`:

- a categorical variable: every level versus the reference level;
- a numeric variable: the per-unit coefficient;
- an interaction such as `genotype:treatment`: every interaction coefficient.

Extra comparisons can be requested with `contrast`. With `at`, a single simple effect is returned instead (for example the treatment effect within one genotype in a `~ genotype * treatment` model). With `test = "lrt"`, one omnibus table is returned and `log2FoldChange/lfcSE` are set to NA because the likelihood-ratio test has no single direction.

Log fold changes are not shrunk.

Value

A list with the fitted `DESeqDataSet` (`dds`), the metadata used, the formula(s), the design matrix, a contrast manifest, and the named result tables (`results`).

Examples

```
ex <- system.file("extdata", "example", package = "rsemflow")
study <- read_rsem_study(file.path(ex, "data"), file.path(ex, "metadata.tsv"))
de <- differential_deseq2(
  study,
  formula = ~ batch + condition,
  term = "condition",
  reference = "condition::Control"
)
head(de$results$condition_Drug_vs_Control)
```

normalize_expression	<i>Normalize or transform expression values</i>
----------------------	---

Description

- vst: DESeq2 size factors, then `varianceStabilizingTransformation(blind = TRUE)`.
- normalized-counts: counts divided by DESeq2 size factors (with the tximport average-length offsets).
- log2-tpm: $\log_2(\text{TPM} + 1)$.

Usage

```
normalize_expression(study, method = c("vst", "normalized-counts", "log2-tpm"))
```

Arguments

study	A rsemflow_study or study path.
method	One of vst, normalized-counts, or log2-tpm.

Value

A gene-by-sample numeric matrix.

read_rsem_study	<i>Import RSEM gene-level results and metadata</i>
-----------------	--

Description

Finds every *.genes.results file in rsem_dir, takes the sample ID from the file name (S01.genes.results -> S01), matches the files to the sample_id column of metadata, and imports the counts with `tximport(type = "rsem")`.

Usage

```
read_rsem_study(  
  rsem_dir,  
  metadata,  
  categorical = character(),  
  numeric = character()  
)
```

Arguments

rsem_dir	Directory containing *.genes.results files.
metadata	Path to TSV or CSV metadata with a sample_id column.
categorical	Optional metadata columns to force to categorical.
numeric	Optional metadata columns to force to numeric.

Details

Import stops when a file lacks the standard RSEM columns, when sample IDs do not match the metadata one-to-one, or when the files do not share an identical gene table (same IDs in the same order, without duplicates).

Metadata columns whose values all parse as numbers become numeric; the rest become factors. Use categorical/numeric to override.

Value

An object of class rsemflow_study.

Examples

```
ex <- system.file("extdata", "example", package = "rsemflow")
study <- read_rsem_study(file.path(ex, "data"), file.path(ex, "metadata.tsv"))
study
```

read_study

Read and write RSEMflow studies

Description

write_study() saves the study as study.rds plus plain TSV copies of the metadata, the expected-count, TPM, and effective-length matrices, a sample-to-file table, and study-info.tsv. read_study() restores the study.rds.

Usage

```
read_study(path)

write_study(study, path)
```

Arguments

path	Study directory or a study.rds path.
study	A rsemflow_study.

Value

read_study() returns an object of class rsemflow_study; write_study() invisibly returns path.

Examples

```
ex <- system.file("extdata", "example", package = "rsemflow")
study <- read_rsem_study(file.path(ex, "data"), file.path(ex, "metadata.tsv"))
out <- file.path(tempdir(), "study")
write_study(study, out)
list.files(out)
identical(read_study(out)$sample_ids, study$sample_ids)
```

run_gsea

*Run preranked GSEA***Description**

Run preranked GSEA

Usage

```
run_gsea(
  differential,
  species = "auto",
  collection = "hallmark",
  subcollection = NULL,
  genesets = NULL,
  rank = "stat",
  min_size = 15L,
  max_size = 500L,
  eps = 0,
  seed = 42L
)
```

Arguments

differential	Differential-expression table or path.
species	auto, human, or mouse.
collection	MSigDB collection alias or raw collection code.
subcollection	Optional MSigDB subcollection.
genesets	Optional named gene-set list or GMT path.
rank	Column used as ranking statistic, default stat.
min_size	Minimum mapped gene-set size.
max_size	Maximum mapped gene-set size.
eps	Passed to fgseaMultilevel.
seed	Random seed for fgseaMultilevel, which estimates p-values by sampling. The same seed gives the same table.

Details

Gene IDs are matched to gene sets by stable Ensembl gene ID. When several rows share a stable ID, the row with the largest absolute ranking statistic is kept.

Unsigned statistics, such as the stat column of an LRT table, are rejected because preranked GSEA needs a direction.

Value

A data frame of GSEA results, sorted by padj and then by decreasing |NES|.

Examples

```
ids <- sprintf("ENSG%011d", 1:100)
de <- data.frame(ensembl_gene_id = ids, stat = seq(5, -5, length.out = 100))
sets <- list(UP = ids[1:20], DOWN = ids[81:100], MIXED = ids[seq(1, 100, 5)])
run_gsea(de, genesets = sets, min_size = 10, max_size = 50)
```

run_gsva	<i>Run GSVA or ssGSEA pathway scoring</i>
----------	---

Description

Scores each gene set in each sample from $\log_2(\text{TPM} + 1)$. Rows sharing a stable Ensembl ID are averaged first. GSVA uses a Gaussian kernel (kcdf = "Gaussian"). Scores are descriptive; no test between groups is run.

Usage

```
run_gsva(
  study,
  species = "auto",
  collection = "hallmark",
  subcollection = NULL,
  genesets = NULL,
  method = c("gsva", "ssgsea"),
  min_size = 10L,
  max_size = 500L
)
```

Arguments

- study A rsemflow_study or study path.
- species auto, human, or mouse.
- collection MSigDB collection alias or raw collection code.
- subcollection Optional MSigDB subcollection.
- genesets Optional named list or GMT file.

method	gsva or ssgsea.
min_size	Minimum gene-set size.
max_size	Maximum gene-set size.

Value

A pathway-by-sample data frame with gene-set metadata columns.

summarize_study	<i>Summarize expression-level QC statistics</i>
-----------------	---

Description

Sample- and gene-level summaries computed from the RSEM expected counts and TPM. A gene is "detected" when its expected count is above zero. These are descriptive numbers only; no sample is flagged or removed.

Usage

```
summarize_study(study)
```

Arguments

study	A rsemflow_study or path to a study.
-------	--------------------------------------

Value

A list with samples and genes data frames.

Examples

```
ex <- system.file("extdata", "example", package = "rsemflow")
study <- read_rsem_study(file.path(ex, "data"), file.path(ex, "metadata.tsv"))
s <- summarize_study(study)
s$samples
```

Index

annotate_ensembl, [2](#)
annotate_gtf, [3](#)

cli_main, [3](#)
compute_correlation, [4](#)
compute_pca, [4](#)

differential_deseq2, [5](#)

normalize_expression, [7](#)

read_rsem_study, [7](#)
read_study, [8](#)
run_gsea, [9](#)
run_gsva, [10](#)

stats::prcomp(), [4](#)
summarize_study, [11](#)

write_study(read_study), [8](#)