Package: ocrRBBR (via r-universe)

May 11, 2026

Type Package

Title Explain Gene Expression with Boolean Rules of Chromatin States

Version 0.1.0

Description Infers Boolean rules among cis-regulatory regions using paired chromatin accessibility and gene expression data at bulk and single-cell levels. Links regulatory regions to target genes, providing insights into gene regulation mechanisms.

License GPL-3

URL https://github.com/CompBioIPM/ocrRBBR

Depends R (>= 4.4)

Imports doParallel, dplyr, foreach, GenomeInfoDb, GenomicRanges, glmnet, Matrix, parallel, rtracklayer, S4Vectors, stats, utils

Config/testthat/edition 3

Encoding UTF-8

LazyData true

LazyDataCompression xz

RoxygenNote 7.3.3

Config/pak/sysreqs make libbz2-dev liblzma-dev libxml2-dev libssl-dev xz-utils zlib1g-dev

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

Date/Publication 2026-02-09 18:16:54 UTC

RemoteUrl https://github.com/compbioipm/ocrrbbr

RemoteRef HEAD

RemoteSha a90a92f46222b838768293c2cee890b2578bdd56

Contents

`ESS` *Estimate Effective Sample Size from Single-Cell RNA-seq Data*

Description

This function estimates the effective sample size (ESS) of single-cell RNA-seq data by accounting for correlation among cells within the same cell type.

Usage

```
ESS(rnaseq_data, cell_type, verbose = FALSE)
```

Arguments

`rnaseq_data`	A numeric matrix of RNA-seq expression values. Rows correspond to genes and columns correspond to cells. Expression values are assumed to be normalized (e.g., `Seurat` `LogNormalize` with `scale.factor = 1e4`).
`cell_type`	A data frame containing cell-type information. Row names must match the column names of `rnaseq_data`. Must include a column named `cell_type` specifying cell types.
`verbose`	Logical. If TRUE, progress messages are printed.

Value

A numeric value representing the effective sample size (ESS), adjusted for within-cell-type correlation.

human_atacseq_data *Human ATAC-seq data*

Description

Human ATAC-seq data

Usage

```
human_atacseq_data
```

Format

An object of class `dgCMatrix` with 43 rows and 9834 columns.

human_meta_data *Human metadata*

Description

Human metadata

Usage

```
human_meta_data
```

Format

An object of class `matrix` (inherits from `array`) with 9834 rows and 3 columns.

human_peaks_gr *Human peaks GRanges*

Description

Human peaks GRanges

Usage

```
human_peaks_gr
```

Format

An object of class `GRanges` of length 108377.

human_rnaseq_data *Human RNA-seq data*

Description

Human RNA-seq data

Usage

```
human_rnaseq_data
```

Format

An object of class dgCMatrix with 2 rows and 9834 columns.

link_peaks_to_tss *Identifies ATAC-seq peaks surrounding each gene*

Description

This function identifies ATAC-seq peaks that are located within a specified window around the transcription start sites (TSS) of genes, and links those peaks to the respective genes.

Usage

```
link_peaks_to_tss(gtf_file, peaks_gr, gene_list = NA, tss_window = NA)
```

Arguments

gtf_file	A character string specifying the path to a GTF file containing gene annotations.
peaks_gr	A GRanges object representing the ATAC-seq peaks with their genomic locations and associated metadata.
gene_list	A character vector of gene names to filter for. Only peaks within the window around TSS of these genes will be considered. Default is NA, in which case all genes in the GTF file will be used.
tss_window	An integer specifying the window size around the TSS. Default is 100000 (±100kb).

Value

A data frame containing the following columns:

peak	The genomic coordinates of the peaks.
peak_id	The unique ID of each peak.
gene_id	The gene ID associated with the peak.
gene_name	The gene name associated with the peak.
transcript_id	The transcript ID associated with the peak.
gene_type	The type of the gene (e.g., protein-coding).
distance	The distance from the peak center to the TSS.

Examples

```
# Load bulk mouse dataset
data(multiome_human_mouse)  # This will load atacseq_data, rnaseq_data, peaks_gr

# Path to the GTF file in the package
gtf_file <- system.file("extdata", "gencode.vM25.annotation.sample.gtf", package = "ocrRBBR")

# Example usage for linking peaks to TSS
linked_peaks <- link_peaks_to_tss(
  gtf_file = gtf_file,
  peaks_gr = mouse_peaks_gr,
  gene_list = c("Rag2"),
  tss_window = 100000
)

# Filter results for a specific gene
linked_peaks_gene <- linked_peaks[linked_peaks$gene_name == "Rag2", ]
```

---

mouse_atacseq_data *Mouse ATAC-seq data*

Description

Mouse ATAC-seq data

Usage

```
mouse_atacseq_data
```

Format

An object of class `data.frame` with 23 rows and 85 columns.

mouse_peaks_gr *Mouse peaks GRanges*

Description

Mouse peaks GRanges

Usage

```
mouse_peaks_gr
```

Format

An object of class GRanges of length 512595.

mouse_rnaseq_data *Mouse RNA-seq data*

Description

Mouse RNA-seq data

Usage

```
mouse_rnaseq_data
```

Format

An object of class data.frame with 2 rows and 85 columns.

ocrRBBR_bulk *Predicts OCR-driven Boolean rules for a gene*

Description

This function predicts Boolean rule sets for a given gene using bulk-level multi-omics datasets, including RNA-seq gene expression and ATAC-seq peak signals in the gene's flanking regions, across samples.

Usage

```
ocrRBBR_bulk(
  rnaseq_data,
  atacseq_data,
  gene_name,
  peak_ids,
  max_feature = NA,
  slope = NA,
  num_cores = NA,
  verbose = FALSE
)
```

Arguments

rnaseq_data	A numeric matrix of RNA-seq expression values. Rows correspond to genes, columns correspond to cell types or samples. **Note:** ocrRBBR was tested using **quantile-normalized RNA-seq data**, but it should also work equally well on **TPM-normalized RNA-seq datasets**, provided the data is appropriately scaled across samples.
atacseq_data	A numeric matrix of ATAC-seq signal intensities. Rows correspond to peaks, columns correspond to cell types or samples. Column names must match those of rnaseq_data. **Note:** Similar to RNA-seq data, ocrRBBR is tested using **quantile-normalized ATAC-seq data** but is expected to work with other normalization methods, as long as the data distributions are comparable across samples.
gene_name	A character string specifying the gene for which to infer Boolean rules.
peak_ids	A vector of peak identifiers corresponding to rows in atacseq_data to be used as candidate regulatory regions for gene_name.
max_feature	An integer specifying the maximum number of input features allowed in a Boolean rule. Default is 3.
slope	The slope parameter for the sigmoid activation function. Default is 10.
num_cores	Number of parallel workers to use for computation. Adjust according to your system. Default is NA (automatic selection).
verbose	Logical. If TRUE, progress messages and a progress bar are shown. Default is FALSE.

Value

A list containing predicted Boolean rules and associated metrics for the input gene.

Examples

```
# Load bulk mouse dataset
data(multiome_human_mouse)  # loads atacseq_data, rnaseq_data, peaks_gr

# Example usage:
peak_ids <- c(278352, 278362, 278381, 278384)
```

```
boolean_rules <- ocrRBBR_bulk(mouse_rnaseq_data, mouse_atacseq_data, "Rag2",
    peak_ids = peak_ids, max_feature = 3, slope = 10, num_cores = 1)
```

`ocrRBBR_single_cell` *Predicts OCR-driven Boolean rules for a gene*

Description

This function predicts Boolean rule sets for a given gene using single-cell multi-omics datasets, including RNA-seq gene expression and ATAC-seq peak signals in the gene's flanking regions, across samples.

Usage

```
ocrRBBR_single_cell(
  rnaseq_data,
  atacseq_data,
  gene_name,
  peak_ids,
  max_feature = NA,
  slope = NA,
  num_cores = NA,
  ESS = NA,
  meta_data,
  verbose = FALSE
)
```

Arguments

`rnaseq_data`	A numeric matrix of RNA-seq expression values. Rows correspond to genes, columns correspond to cells or samples. RNA-seq values are assumed to be normalized using Seurat's `LogNormalize` method with a scale factor of 10,000: `NormalizeData(seurat_obj, normalization.method = "LogNormalize", scale.factor = 1e4)`.
`atacseq_data`	A numeric matrix of ATAC-seq signal intensities. Rows correspond to peaks, columns correspond to cells or samples. ATAC-seq counts are assumed to be normalized by the `ReadsInTSS` metric on a per-cell basis, where raw peak counts are divided by the number of Tn5 insertions falling within transcription start site (TSS) regions for each cell. This normalization corrects for differences in sequencing depth and chromatin accessibility signal across cells. `ReadsInTSS` values are typically obtained from `ArchR` and applied as column-wise scaling factors to the ATAC-seq count matrix prior to downstream analysis.
`gene_name`	A character string specifying the gene for which to infer Boolean rules.
`peak_ids`	A vector of peak identifiers corresponding to rows in `atacseq_data` to be used as candidate regulatory regions for `gene_name`.

max_feature	An integer specifying the maximum number of input features allowed in a Boolean rule. Default is 3.
slope	The slope parameter for the sigmoid activation function. Default is 10.
num_cores	Number of parallel workers to use for computation. Adjust according to your system. Default is NA.
ESS	Effective sample size of the single-cell data after accounting for noise and cell-to-cell correlation.
meta_data	A numeric matrix or data.frame containing additional per-cell covariates (rows = cells, columns = covariates), such as: **nCount_RNA**: The total number of RNA molecules (unique molecular identifiers, UMIs) detected per cell, calculated as the sum of UMI counts across all genes. **nFeature_RNA**: The number of genes detected per cell (genes with at least one UMI). **Mitochondrial percentage**: The percentage of reads that map to mitochondrial genes, which can be used to assess the quality of the sample. This information is typically stored as columns in the meta_data object, which is associated with each cell in the dataset.
verbose	Logical. If TRUE, progress messages and a progress bar are shown. Default is FALSE.

Value

A list containing predicted Boolean rules and associated metrics for the input gene.

Examples

```
# Load single-cell human dataset
data(multiome_human_mouse)  # loads atacseq_data, rnaseq_data, peaks_gr, meta_data

# Example usage:
peak_ids <- c(83456, 83458, 83460)

boolean_rules <- ocrRBBR_single_cell(human_rnaseq_data, human_atacseq_data,
       "CD74", peak_ids = peak_ids, max_feature = 3, slope = 6,
        num_cores = 1, ESS = 261, meta_data = human_meta_data)
```

Index