Package 'mist'

January 20, 2025

Title Differential Methylation Analysis for scDNAm Data

Version 0.99.17

Depends R (>= 4.5.0)

Description mist (Methylation Inference for Single-cell along Trajectory) is a hierarchical Bayesian framework for modeling DNA methylation trajectories and performing differential methylation
(DM) analysis in single-cell DNA methylation (scDNAm) data. It estimates developmental-stage-specific
variations, identifies genomic features with drastic changes along pseudotime, and, for two phenotypic groups, detects features with distinct temporal methylation patterns. mist uses Gibbs sampling to estimate parameters for temporal changes and stage-specific variations.

License MIT + file LICENSE

Encoding UTF-8

URL https://https://github.com/dxd429/mist

BugReports https://https://github.com/dxd429/mist/issues

biocViews Epigenetics, DifferentialMethylation, DNAMethylation, SingleCell, Software

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Imports BiocParallel, MCMCpack, Matrix, S4Vectors, methods, rtracklayer, car, mvtnorm, SummarizedExperiment, SingleCellExperiment, BiocGenerics, stats, rlang

Suggests knitr, rmarkdown, RUnit, ggplot2, BiocStyle

VignetteBuilder knitr

git_url https://git.bioconductor.org/packages/mist

git_branch devel

git_last_commit 9d19bea

git_last_commit_date 2025-01-14

Repository Bioconductor 3.21

Date/Publication 2025-01-19

Author Daoyu Duan [aut, cre] (ORCID: <https://orcid.org/0000-0002-3147-2006>)

Maintainer Daoyu Duan <dxd429@case.edu>

Contents

`dmSingle` *Differential methylation evaluation over time for scDNA-seq data under the 1-group scenario.*

Description

This function performs DM analysis to identify genomic features showing drastic changes along pseudotime, by modeling the methylation changes along pseudotime and estimating the developmental-stage-specific biological variations.

Usage

```
dmSingle(Dat_sce, BPPARAM = SnowParam())
```

Arguments

`Dat_sce`	The updated sce object with A numeric matrix of estimated parameters for all genomic features in the rowData, including: • β_0 to β_4: Estimated coefficients for the polynomial of degree 4. • σ_1^2 to σ_4^2: Estimated variances for each stage along the pseudotime.
`BPPARAM`	A `BiocParallelParam` object specifying the parallel backend for computations, as used in `bplapply()`. Defaults to `SnowParam()` for cluster-based parallel processing.

Value

The updated sce object with A named numeric vector where each value corresponds to a genomic feature (e.g., a gene) in the rowData. The values represent the minimum area between the fitted curve of scDNA methylation levels along pseudotime and a constant horizontal line for each feature. This metric measures the magnitude of methylation level changes along pseudotime. The values are sorted in descending order, with larger values indicating more drastic changes.

Examples

```
library(mist)
data <- readRDS(system.file("extdata", "small_sampleData_sce.rds", package = "mist"))
Dat_sce_new <- estiParamSingle(
    Dat_sce = data,
    Dat_name = "Methy_level_group1",
    ptime_name = "pseudotime"
)
dm_sce <- dmSingle(Dat_sce_new)
```

`dmTwoGroups` *Differential methylation evaluation over time for scDNA-seq data under the 2-group scenario.*

Description

This function performs DM analysis to identify genomic features showing drastic changes between two groups along pseudotime, by modeling the methylation changes and comparing the fitted curves for each group.

Usage

```
dmTwoGroups(Dat_sce, BPPARAM = SnowParam())
```

Arguments

`Dat_sce`	The updated sce object with 2 numeric matrices in the rowData: • `mist_pars_group1`: A numeric matrix of estimated parameters for all genomic features in group 1 (generated by `estiParamTwoGroups`). • `mist_pars_group2`: A numeric matrix of estimated parameters for all genomic features in group 2.
`BPPARAM`	A `BiocParallelParam` object specifying the parallel backend for computations, as used in `bplapply()`. Defaults to `SnowParam()` for cluster-based parallel processing.

Value

The updated sce object with a named numeric vector where each value corresponds to a genomic feature (e.g., a gene). The values represent the integral of the differences between the fitted curves of scDNA methylation levels for the two groups. The values are sorted in descending order, with larger values indicating more drastic differences between the groups.

Examples

```
library(mist)
data <- readRDS(system.file("extdata", "small_sampleData_sce.rds", package = "mist"))
Dat_sce_new <- estiParamTwo(
    Dat_sce = data,
    Dat_name_g1 = "Methy_level_group1",
    Dat_name_g2 = "Methy_level_group2",
    ptime_name_g1 = "pseudotime",
    ptime_name_g2 = "pseudotime_g2"
)
dm_sce_two <- dmTwoGroups(Dat_sce_new)
```

`estiParamSingle` *Parameter Estimation for Single-Group*

Description

This function performs the Gibbs sampling procedure based on hierarchical Bayesian modeling to produce the parameters required for differential methylation analysis.

Usage

```
estiParamSingle(
  Dat_sce,
  Dat_name,
  ptime_name,
  BPPARAM = SnowParam(),
  verbose = TRUE
)
```

Arguments

`Dat_sce`	A `SingleCellExperiment` object containing the single-cell DNA methylation level. Methylation levels should be stored as an assay, with genomic feature (gene) names in rownames and cells in colnames.
`Dat_name`	A character string specifying the name of the assay to extract the methylation level data.
`ptime_name`	A character string specifying the name of the column in `colData` containing the pseudotime vector.
`BPPARAM`	A `BiocParallelParam` object specifying the parallel backend for computations, as used in `bplapply()`. Defaults to `SnowParam()` for cluster-based parallel processing.
`verbose`	A logical value indicating whether to print progress messages to the console. Defaults to `TRUE`. Set to `FALSE` to suppress messages.

Value

The updated sce object with A numeric matrix of estimated parameters for all genomic features in the rowData, including:

- β_0 to β_4: Estimated coefficients for the polynomial of degree 4.
- σ_1^2 to σ_4^2: Estimated variances for each stage along the pseudotime.

Examples

```
library(SingleCellExperiment)
data <- readRDS(system.file("extdata", "small_sampleData_sce.rds", package = "mist"))
Dat_sce_new <- estiParamSingle(
    Dat_sce = data,
    Dat_name = "Methy_level_group1",
    ptime_name = "pseudotime"
)
```

`estiParamTwo` *Parameter Estimation for Two-Group Scenario*

Description

This function performs the Gibbs sampling procedure based on hierarchical Bayesian modeling, separately for two groups, to produce the parameters required for differential methylation analysis.

Usage

```
estiParamTwo(
  Dat_sce,
  Dat_name_g1,
  Dat_name_g2,
  ptime_name_g1,
  ptime_name_g2,
  BPPARAM = SnowParam(),
  verbose = TRUE
)
```

Arguments

`Dat_sce`	A `SingleCellExperiment` object containing the single-cell DNA methylation level. Methylation levels should be stored as assays, with genomic feature (gene) names in rownames and cells in colnames.
`Dat_name_g1`	A character string specifying the name of the assay to extract the group 1 methylation level data.
`Dat_name_g2`	A character string specifying the name of the assay to extract the group 2 methylation level data.

ptime_name_g1	A character string specifying the name of the column in colData for the group 1 pseudotime vector.
ptime_name_g2	A character string specifying the name of the column in colData for the group 2 pseudotime vector.
BPPARAM	A BiocParallelParam object specifying the parallel backend for computations, as used in bplapply(). Defaults to SnowParam() for cluster-based parallel processing.
verbose	A logical value indicating whether to print progress messages to the console. Defaults to TRUE. Set to FALSE to suppress messages.

Value

The updated sce object with two matrices in the rowData, one for each group, where each matrix contains estimated parameters for all genomic features, including:

- β_0 to β_4: Estimated coefficients for the polynomial of degree 4.
- σ_1^2 to σ_4^2: Estimated variances for each stage along the pseudotime.

Examples

```
library(SingleCellExperiment)
data <- readRDS(system.file("extdata", "small_sampleData_sce.rds", package = "mist"))
Dat_sce_new <- estiParamTwo(
    Dat_sce = data,
    Dat_name_g1 = "Methy_level_group1",
    Dat_name_g2 = "Methy_level_group2",
    ptime_name_g1 = "pseudotime",
    ptime_name_g2 = "pseudotime_g2"
)
```

---

plotGene *Plot Methylation Levels vs. Pseudotime for a Specific Gene*

Description

Generates a scatter plot of methylation levels vs. pseudotime for a specific gene, overlaid with the fitted curve based on the estimated coefficients.

Usage

```
plotGene(Dat_sce, Dat_name, ptime_name, gene_name)
```

Arguments

Dat_sce	The updated sce object with A numeric matrix of estimated parameters for all genomic features in the rowData, including: • β_0 to β_4: Estimated coefficients for the polynomial of degree 4. • σ_1^2 to σ_4^2: Estimated variances for each stage along the pseudotime.
Dat_name	A character string specifying the name of the assay to extract the methylation data.
ptime_name	A character string specifying the name of the colData to extract the pseudotime vector.
gene_name	A character string specifying the gene name to plot.

Value

A ggplot2 scatter plot with an overlayed fitted curve.

Examples

```
library(SingleCellExperiment)
library(ggplot2)
data <- readRDS(system.file("extdata", "small_sampleData_sce.rds", package = "mist"))
Dat_sce_new <- estiParamSingle(
    Dat_sce = data,
    Dat_name = "Methy_level_group1",
    ptime_name = "pseudotime"
)
plotGene(Dat_sce = Dat_sce_new,
Dat_name = "Methy_level_group1",
ptime_name = "pseudotime",
gene_name = "ENSMUSG00000000037")
```

---

sampleData_sce *Sample Single-Cell Data for Testing and Illustration*

Description

A single-cell dataset for testing and illustration purposes, derived from the GEO dataset GSE121708. The data contains DNA methylation information from mouse gastrulation experiments at the single-cell level.

Format

A `SingleCellExperiment` object containing:

- Metadata about cells and genes.
- DNA methylation levels across a subset of genes and cells.

Details

- **Data Processing Steps:**
  1. Annotation: Data was annotated using the mm10 reference genome.
  2. Data Cleaning: Genes expressed in less than 10% of cells were removed, resulting in a scDNAm dataset containing 986 cells and 18,220 genes.
  3. Pseudotime Inference: Pseudotime was inferred using the default settings in Monocle3.
  4. Subsampling:
     - For illustration purposes (e.g., vignette generation), a random sample of 100 genes was selected to create `sampleData_sce.rds`.

Source

The data originates from the GEO dataset GSE121708, as described in:

1. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 2019 Dec;576(7787):487-491. PMID: 31827285.
2. Kapourani CA, Argelaguet R, Sanguinetti G, Vallejos CA. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. Genome Biol 2021 Apr 20;22(1):114. PMID: 33879195.

Examples

```
file_path <- system.file("extdata", "sampleData_sce.rds", package = "mist")
Dat_sce <- readRDS(file_path)
```

---

small_sampleData_sce *Small Sample Single-Cell Data for Testing and Illustration*

Description

A small single-cell dataset for testing and illustration purposes, derived from the GEO dataset GSE121708. The data contains DNA methylation information from mouse gastrulation experiments at the single-cell level.

Format

A `SingleCellExperiment` object containing:

- Metadata about cells and genes.
- DNA methylation levels across a subset of genes and cells.

Details

- **Data Processing Steps:**
  1. Annotation: Data was annotated using the mm10 reference genome.
  2. Data Cleaning: Genes expressed in less than 10% of cells were removed, resulting in a scDNAm dataset containing 986 cells and 18,220 genes.
  3. Pseudotime Inference: Pseudotime was inferred using the default settings in Monocle3.
  4. Subsampling:
     - For illustration purposes (e.g., vignette generation), a random sample of 100 genes was selected to create `sampleData_sce.rds`.
     - For unit tests and function help pages, a random sample of 5 genes was selected to create `small_sampleData_sce.rds`.

Source

The data originates from the GEO dataset GSE121708, as described in:

1. Argelaguet R, Clark SJ, Mohammed H, Stapel LC, et al. Multi-omics profiling of mouse gastrulation at single-cell resolution. Nature 2019 Dec;576(7787):487-491. PMID: 31827285.
2. Kapourani CA, Argelaguet R, Sanguinetti G, Vallejos CA. scMET: Bayesian modeling of DNA methylation heterogeneity at single-cell resolution. Genome Biol 2021 Apr 20;22(1):114. PMID: 33879195.

Examples

```
file_path <- system.file("extdata", "small_sampleData_sce.rds", package = "mist")
Dat_sce <- readRDS(file_path)
```

Index