Package 'idiffomix'

January 13, 2025

Type Package

Title Integrated Differential Analysis of Multi Omics Data using a Joint Mixture Model

Version 1.0.0

Maintainer Koyel Majumdar <koyelmajumdar.phdresearch@gmail.com>

Description A joint mixture model has been developed by Majumdar et al. (2025) <doi:10.48550/arXiv.2412.17511> that integrates information from gene expression data and methylation data at the modelling stage to capture their inherent dependency structure, enabling simultaneous identification of differentially methylated cytosine-guanine dinucleotide (CpG) sites and differentially expressed genes. The model leverages a joint likelihood function that accounts for the nested structure in the data, with parameter estimation performed using an expectation-maximisation algorithm.

License GPL-3

Depends R (>= 3.5.0)

Imports foreach, doParallel, parallel, mclust, stats, utils, edgeR, magrittr, ggplot2, scales, tidyr, dplyr, reshape2, gridExtra, grid, tidyselect, cowplot

Encoding UTF-8

NeedsCompilation yes

LazyData true

LazyDataCompression xz

RoxygenNote 7.3.2

Suggests rmarkdown, knitr

VignetteBuilder knitr

Author Koyel Majumdar [cre, aut], Isobel Claire Gorley [aut], Thomas Brendan Murphy [aut], Florence Jaffrezic [aut], Andrea Rau [aut]

Repository CRAN

Date/Publication 2025-01-13 17:20:02 UTC

Contents

`data_transformation` *The function to filter, normalize and transform RNA-Seq and methylation data.*

Description

The raw RNA-Seq and methylation data needs to be filtered, normalized and transformed before applying the idiffomix method.

Usage

```
data_transformation(seq_data, meth_data, gene_chr, N = 5)
```

Arguments

`seq_data`	A dataframe of dimension $G \times (N + 1)$ containing raw RNA-Seq data for all G genes and N patients
`meth_data`	A dataframe of dimension $C \times (N + 2)$ containing beta methylation values for all C CpG sites and N patients along with the associated genes for each CpG site.
`gene_chr`	A dataframe containing the genes and their corresponding chromosome number.
`N`	Number of patients

Details

The RNA-Seq data consisted of raw counts depicting the gene expression levels. To ensure data quality, only genes whose sum of expression counts across both biological conditions was > 5 are retained. The data were normalized to account for differences in library sizes. The normalized count data were used to obtain CPM values which were further log-transformed to obtain log-CPM values. Given the paired design of the motivating setting, the log-fold changes between the tumour and benign samples were calculated for each gene in every patient and used in the subsequent analyses. For the methylation array data, the beta values at the CpG sites are logit transformed to M-values. Similar to the RNA-Seq data, given the paired design, the difference in M-values between tumour and benign samples were calculated for each CpG site in every patient and used in the subsequent analyses.

Value

The function returns a list with two dataframes containing the transformed gene expression and methylation array data:

- seq_transformed - A dataframe containing the log-fold change for gene expression data.
- meth_transformed - A dataframe containing the differences in M-values for methylation data.

Examples

```
N <- 4
data_output = data_transformation(seq_data=gene_expression_data,
                                  meth_data=methylation_data,
                                  gene_chr=gene_chromosome_data,
                                  N=N)
```

---

gene_chromosome_data *Data containing chromosome information and the genes located on them.*

Description

A dataset containing the chromosome information of the gene expression and methylation array data to be analysed.

Usage

```
data(gene_chromosome_data)
```

Format

A data frame with 20 rows and 2 columns.

- CHR: The chromosome containing the gene.
- Gene: The gene located on the chromosome.

Details

The dataset contains the information of chromosomes 1 AND 2 and the genes located on them.

See Also

gene_expression_data

methylation_data

gene_expression_data *Gene expression data for patients suffering from breast cancer*

Description

A dataset containing simulated RNA-Seq data for G genes located on chromosomes 1 and 2, from $R = 2$ sample types, from $N = 4$ patients. The sample types are assumed to be benign and tumour tissues.

Usage

```
data(gene_expression_data)
```

Format

A data frame with 20 rows and 9 columns. The data contain no missing values.

- Gene: The gene name.
- Patient1_GX1: Expression values from benign tissue from patient 1.
- Patient2_GX1: Expression values from benign tissue from patient 2.
- Patient3_GX1: Expression values from benign tissue from patient 3.
- Patient4_GX1: Expression values from benign tissue from patient 4.
- Patient1_GX2: Expression values from tumour tissue from patient 1.
- Patient2_GX2: Expression values from tumour tissue from patient 2.
- Patient3_GX2: Expression values from tumour tissue from patient 3.
- Patient4_GX2: Expression values from tumour tissue from patient 4.

Details

The simulated raw RNA-Seq data for genes located on the chromosomes 1 and 2 needs to be filtered, normalized and transformed before applying idiffomix.

See Also

gene_chromosome_data

methylation_data

idiffomix *The idiffomix function*

Description

Integrated differential analysis of multi-omics data using a joint mixture model

Usage

```
idiffomix(
  seq_data,
  meth_data,
  gene_chr,
  N,
  K = 3,
  L = 3,
  probs = c(0.1, 0.9),
  parallel_process = FALSE
)
```

Arguments

seq_data	A dataframe of dimension $G \times (N + 1)$ containing log-fold change values for all G genes and N patients
meth_data	A dataframe of dimension $C \times (N + 2)$ containing M-value differences between the two biological conditions for all C CpG sites and N patients along with the associated genes for each CpG site.
gene_chr	A dataframe containing the genes and their corresponding chromosome number.
N	Number of patients in the study.
K	Number of clusters for expression data (default = 3; E-,E0,E+).
L	Number of clusters for methylation data (default = 3; M-,M0,M+).
probs	Quantile probabilities for initialization (default = c(0.1,0.9)).
parallel_process	The "TRUE" option results in parallel processing of the models for increased computational efficiency. The default option has been set as "FALSE" due to package testing limitations.

Details

This is a function to fit the joint mixture model to the transformed and filtered gene expression and methylation data.

Value

The function returns an object of the idiffomix class which contains the following values:

- tau - The proportion of genes belonging to K clusters.
- pi - A matrix containing the probability of a CpG site belonging to cluster *l*, given its associated associated gene belongs to cluster *k*.
- mu - The mean for each component of the gene expression data. If there is more than one component, this is a matrix whose *k*th column is the mean of the *k*th component of the mixture model.
- sigma - The variance for each component of the gene expression data.
- lambda - The mean for each component of the methylation data. If there is more than one component, this is a matrix whose *l*th column is the mean of the *l*th component of the mixture model.
- rho - The variance for each component of the methylation data.
- N - The number of patients analysed using the beta mixture models.
- R - The number of sample types analysed using the beta mixture models.
- U - A dataframe containing the posterior probabilities of genes belonging to the *K* clusters.
- V - A dataframe containing the posterior probabilities of CpG sites belonging to the *L* clusters.
- seq_classification - A dataframe containing the log-fold change for gene expression data and their classification corresponding to U.
- meth_classification - A dataframe containing the differences in M-values for methylation data and their classification corresponding to V.

Examples

```
N <- 4
data_transformed = data_transformation(seq_data=gene_expression_data,
                                       meth_data=methylation_data,
                                       gene_chr=gene_chromosome_data,
                                       N=N)
data_output = idiffomix(seq_data=data_transformed$seq_transformed,
                        meth_data=data_transformed$meth_transformed,
                        gene_chr=gene_chromosome_data,
                        N=N, K=3, L=3, probs=c(0.25,0.75),
                        parallel_process = FALSE)
```

---

methylation_data *Methylation array data for patients suffering from breast cancer*

Description

A dataset containing simulated methylation array data for C CpG sites associated to G genes, from $R = 2$ sample types, collected from $N = 4$ patients.

Usage

```
data(methylation_data)
```

Format

A data frame with 205 rows and 10 columns.

- Gene: The gene name.
- CpG: The CpG site associated to the gene.
- Patient1_M1: Methylation values from benign tissue from patient 1 for the corresponding CpG site.
- Patient2_M1: Methylation values from benign tissue from patient 2 for the corresponding CpG site.
- Patient3_M1: Methylation values from benign tissue from patient 3 for the corresponding CpG site.
- Patient4_M1: Methylation values from benign tissue from patient 4 for the corresponding CpG site.
- Patient1_M2: Methylation values from tumour tissue from patient 1 for the corresponding CpG site.
- Patient2_M2: Methylation values from tumour tissue from patient 2 for the corresponding CpG site.
- Patient3_M2: Methylation values from tumour tissue from patient 3 for the corresponding CpG site.
- Patient4_M2: Methylation values from tumour tissue from patient 4 for the corresponding CpG site.

Details

The methylation array data is assumed to be from benign and tumour tissues. The methylation data comprised of beta-valued methylation levels. The data needs to be transformed before applying idiffomix.

See Also

gene_expression_data

gene_chromosome_data

`plot.idiffomix` *Plots for visualizing the idiffomix class object*

Description

Visualise a `idiffomix` clustering solution by plotting the conditional probabilities estimated for genes and CpG sites (A) per chromosome and (B) for a gene and its corresponding CpG sites.

Usage

```
## S3 method for class 'idiffomix'
plot(
  x,
  what = "chromosome",
  CHR = 1,
  Gene = NULL,
  K = 3,
  L = 3,
  gene_cluster_name = c("E-", "E0", "E+"),
  cpg_cluster_name = c("M-", "M0", "M+"),
  ...
)
```

Arguments

`x`	A `idiffomix` object.
`what`	The different plots that can be obtained are either "chromosome" or "gene" (default = "chromosome").
`CHR`	The chromosome number to be visualized (default = 1).
`Gene`	The name of the gene to be visualized (default = NULL).
`K`	Number of clusters for expression data (default = 3; E-,E0,E+).
`L`	Number of clusters for methylation data (default = 3; M-,M0,M+).
`gene_cluster_name`	The names to be given to the clusters for identification (default = c("E-","E0","E+")).
`cpg_cluster_name`	The names to be given to the clusters for identification (default = c("M-","M0","M+")).
`...`	Other graphics parameters.

Details

The function displays two plots. The first plot displays the conditional probabilities estimated when the joint mixture model is applied to data from a chromosome. Panel A displays the probability of a gene in the chromosome belonging to each of the K clusters. Panel B details the estimated matrix pi of conditional probabilities of a CpG site's cluster membership, given its gene's cluster. Panel C details the conditional probabilities of a gene belonging to cluster k given a single CpG

site associated with the gene belongs to cluster l, computed using Bayes' theorem, given tau and pi. The second plot displays the log-fold changes and differences in M-values and the estimated posterior probability of the gene belonging to the K clusters. Panel A shows the log-fold change and difference in M-values for the given gene and its associated CpG sites while Panel B shows the posterior probabilities of cluster membership for the gene under idiffomix.

Value

This function displays the following plots as requested by the user:

- chromosome plot - Plot showing the conditional probabilities estimated when the joint mixture model is applied to data from a chromosome.
- gene plot - Plot showing the log-fold changes and differences in M-values and the estimated posterior probability of the gene belonging to the K clusters.

See Also

idiffomix

Examples

```
N <- 4
data_transformed = data_transformation(seq_data=gene_expression_data,
                                  meth_data=methylation_data,
                                  gene_chr=gene_chromosome_data,
                                  N=N)
data_output = idiffomix(seq_data=data_transformed$seq_transformed,
                        meth_data=data_transformed$meth_transformed,
                        gene_chr=gene_chromosome_data,
                        N=N, K=3, L=3, probs=c(0.25,0.75),
                        parallel_process = FALSE)
plot(data_output,what="chromosome",CHR=1, Gene=NULL,K=3,L=3,
     gene_cluster_name=c( "E-","E0","E+"),
     cpg_cluster_name=c( "M-","M0","M+"),
     title=NULL)
```

Index