Package ‘SIMD’

May 30, 2024

Type Package

Title Statistical Inferences with MeDIP-seq Data (SIMD) to infer the methylation level for each CpG site

Version 1.22.0

Author Yan Zhou

Maintainer Jiadi Zhu <2160090406@email.szu.edu.cn>

Description This package provides a inferential analysis method for detecting differentially expressed CpG sites in MeDIP-seq data. It uses statistical framework and EM algorithm, to identify differentially expressed CpG sites. The methods on this package are described in the article 'Methylation-level Inferences and Detection of Differential Methylation with Medip-seq Data' by Yan Zhou, Jiadi Zhu, Mingtao Zhao, Baoxue Zhang, Chunfu Jiang and Xiyan Yang (2018, pending publication).

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.5.0)

Imports edgeR, statmod, methylMnM, stats, utils

Suggests BiocStyle, knitr,rmarkdown

biocViews ImmunoOncology, DifferentialMethylation,SingleCell, DifferentialExpression

VignetteBuilder knitr

RoxygenNote 6.0.1

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

git_branch RELEASE_3_19

git_last_commit ad56a8a

git_last_commit_date 2024-04-30

Repository Bioconductor 3.19

Date/Publication 2024-05-29
Contents

SIMD-package ........................................... 2
all_CpGsite_bin_chr18 .................................. 3
classifypvalue ........................................ 3
EM2_H1ESB1_MeDIP_sigleCpG ......................... 4
EM2_H1ESB2_MeDIP_sigleCpG ......................... 5
EMalgorithm ............................................ 5
emalgth .................................................. 6
emalgth1 ................................................ 7
EMtest .................................................... 7
EM_H1ESB1_MeDIP_sigleCpG ......................... 9
H1ESB1_MRE_sigleCpG .................................. 9
H1ESB2_MRE_sigleCpG .................................. 10
probBinom ............................................. 11
three_mre_cpg ......................................... 11

Index ..................................................... 13

SIMD-package ........................................... 2
A method to infer the methylation expression level for each CpG sites.

Description

SIMD is a package to infer the methylation expression level for each CpG sites. The main idea of SIMD is that by using statistical inference to with Medip-seq data method to infer the methylation level.

Author(s)

Zhou Yan Maintainer: Zhou Yan <zhouy1016@szu.edu.cn>

References

**all_CpGsite_bin_chr18**

*A simulation dataset of CpG sites.*

**Description**

This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the CpG sites.

**Usage**

`all_CpGsite_bin_chr18`

**Format**

A data.frame containing 2000 CpG sites.

**Source**


**References**


classifypvalue  

*calculate P-value in code EMtest.*

**Description**

calculate P-value in code EMtest.

**Usage**

`classifypvalue(type1, type2, type3, type4, smchr1ring1, smchr1ring2, smchr1ring3, smchr1ring4, p, typelength, smchr1ringlength, pvalue = rep(0, length(smchr1ring1)))`
EM2_H1ESB1_MeDIP_sigeCpG

Arguments

- **type1**: The first column of the first matrix.
- **type2**: The second column of the first matrix.
- **type3**: The third column of the first matrix.
- **type4**: The fourth column of the first matrix.
- **sm1chr1**: The first column of the second matrix.
- **sm1chr2**: The second column of the second matrix.
- **sm1chr3**: The third column of the second matrix.
- **sm1chr4**: The fourth column of the second matrix.
- **p**: P-value.
- **type1length**: The number of rows of the first matrix.
- **sm1chr1length**: The number of rows of the second matrix.
- **pvalue**: A vector, the length equals to the number of rows of the second matrix.

Value

The probability.

---

EM2_H1ESB1_MeDIP_sigeCpG

* A simulation dataset of MeDIP CpG sites.

Description

This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MeDIP CpG sites.

Usage

EM2_H1ESB1_MeDIP_sigeCpG

Format

A data.frame containing 2000 MeDIP CpG sites.

Source


References

**EM2_H1ESB2_MeDIP_sigeCpG**

*A simulation dataset of MeDIP CpG sites.*

---

**Description**

This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MeDIP CpG sites.

**Usage**

**EM2_H1ESB2_MeDIP_sigeCpG**

**Format**

A data.frame containing 2000 MeDIP CpG sites.

**Source**


**References**


---

**EMalgorithm**

*EM algorithm to infer CpG sites.*

---

**Description**

Using EM algorithm to infer the real number of CpG sites.

**Usage**

**EMalgorithm(cpgsitefile, allcpgfile, category = "1", writefile = NULL, reportfile = NULL)**

**Arguments**

- `cpgsitefile`: The path of file to store CpG site.
- `allcpgfile`: The file to store CpG sites.
- `category`: Default to "1".
- `writefile`: The path of output results. (If writefile=NULL, there will return the results back to main program.)
- `reportfile`: The path of output results.
emalgth

Calculate the probability on condition that the sums equal to 1.

Description
Calculate the probability on condition that only a single CpG contributes to a short read.

Usage
emalgth(X)

Arguments
X  A matrix about X, the elements in X takes values on 0,1 and satisfy the sums of each row equal to 1.

Value
y1  The probability when sums equal to 1.

Examples
set.seed(123)
d <- matrix(0, nrow=200, ncol=50)
random_num <- sample(1:50, 200, replace=TRUE)
for(i in 1:nrow(d)){
  d[i,random_num[i]]<-1
}
result <- emalgth(d)
head(result)
Calculate the probability on condition that the sums more than 1.

**Description**

Calculate the probability on condition that at least a CpG contributes to a short read.

**Usage**

```r
calculate_prob <- function(X) {
  # Implementation
}
```

**Arguments**

- `X`: A matrix about X, the elements in X takes values on 0,1 and satisfy the sums of each row more than 1.

**Value**

- `y1`: The probability when sums more than 1.

**Examples**

```r
set.seed(123)
X <- matrix(0, nrow=200, ncol=50)
random_num <- sample(1:10, 200, replace=TRUE)
for(i in 1:nrow(X)){
  temp <- sample(1:50, random_num[i], replace=FALSE)
  X[i,temp] <- 1
}
result <- calculate_prob(X)
head(result)
```

**EMtest**

Inferring the methylation expression level of single sites.

**Description**

Using statistical framework and EM algorithm to infer the methylation expression level of single sites.

**Usage**

```r
EMtest(datafile = NULL, chrstring = NULL, cpgfile, mrecpgfile = NULL,
       writefile = NULL, reportfile = NULL, mrratio = 3/7, psd = 2,
       mkadded = 1, f = 1)
```
Arguments

**datafile**  
The files of sample. (datafile should be cbind(data1, data2, data3, data4), where data1 and data2 are Medip-seq data, data3 and data4 are MRE-seq data).

**chrstring**  
The chromosome should be test.

**cpgfile**  
The file of all CpG number.

**mrecpgfile**  
The file of MRE-CpG number (If NULL, mrecpgfile will equal to cpgfile).

**writefile**  
The path of file of output result. (If writefile=NULL, there will return the results back to main program)

**reportfile**  
The path of output results of the number of bin, total reads before processing and total reads after processing.

**mreratio**  
The ratio of total unmethylation level with total methylation level (Defaulted mreratio is 3/7).

**psd**  
The parameters of pseudo count, which pseudo count added to Medip-seq and MRE-seq count.

**mkadded**  
Added to all CpG and MRE CpG (We set psd=2 and mkadded=1 as defaulted for robust).

**f**  
Adjustment weight, default to 1.

Value

values or file The output file "writefile" will own eleven columns, that is, "chr", "chrSt", "chrEnd", "Medip1", "Medip2", "MRE1", "MRE2", "cg", "mrecg", "pvalue" and "Ts". We also output a report file which will include parameters "s1/s2", "s3/s4", "N1", "N2", "N3", "N4", "c1", "c2", "Number of windows" and "Spend time".

Examples

```r
data(example_data)
data1 <- EM2_H1ESB1_MeDIP_sigeCpG
data2 <- EM2_H1ESB2_MeDIP_sigeCpG
data3 <- H1ESB1_MRE_sigeCpG
data4 <- H1ESB2_MRE_sigeCpG
datafile <- cbind(data1, data2, data3, data4)
callcpg <- all_CpGsite_bin_chr18
mrecpg <- three_mre_cpg
dirwrite <- paste(setwd(getwd()), "/", sep="")
writefile <- paste(dirwrite, "pval_EM_H1ESB1_H1ESB21.bed", sep="")
reportfile <- paste(dirwrite, "report_pvalH1ESB1_H1ESB21.bed", sep="")
EMtest(datafile=datafile, chrstring=NULL, cpgfile=callcpg, mrecpgfile=mrecpg, writefile=writefile, reportfile=reportfile,
       mreratio=3/7, psd=2, mkadded=1, f=1)
```
EM_H1ESB1_MeDIP_sigleCpG

A simulation dataset of MeDIP CpG sites.

Description
This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MeDIP CpG sites.

Usage
EM_H1ESB1_MeDIP_sigleCpG

Format
A data.frame containing 2000 MeDIP CpG sites.

Source

References

H1ESB1_MRE_sigleCpG

A simulation dataset of MRE CpG sites.

Description
This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MRE CpG sites.

Usage
H1ESB1_MRE_sigleCpG

Format
A data.frame containing 2000 MRE CpG sites.
Source


References


H1ESB2_MRE_sigeCpG  A simulation dataset of MRE CpG sites.

Description

This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MRE CpG sites.

Usage

H1ESB2_MRE_sigeCpG

Format

A data.frame containing 2000 MRE CpG sites.

Source


References

**probBinom**

*Compute P-values for Medip-seq and MRE-seq data.*

**Description**

Compute P-values.

**Usage**

```r
probBinom(t, size1, size2, c1, c2)
```

**Arguments**

- `t` The real value for random variable according to dataset.
- `size1` The sum of Medip-seq real reads of the each CpG site for control and treatment sample.
- `size2` The sum of MRE-seq real reads of the each CpG site for control and treatment sample.
- `c1` The scaling factor for MeDip-seq data.
- `c2` The scaling factor for MRE-seq data.

**Value**

- `p` The P-values for testing the methylation expression levels for each CpG sites.

**Examples**

```r
set.seed(1234)
t <- 0.1
size1 <- sample(1:1000, 1, replace=TRUE)
size2 <- sample(1:1000, 1, replace=TRUE)
c1 <- 1
c2 <- 2
result <- probBinom(t, size1, size2, c1, c2)
```

---

**three_mre_cpg**

*A simulation dataset of MRE CpG sites.*

**Description**

This data set gives 2000 CpG sites which include the chromosome of the region, the start and the stop position of the MRE CpG sites.

**Usage**

```r
three_mre_cpg
```
Format
A data.frame containing 2000 MRE CpG sites.

Source

References
Index

* datasets
  all_CpGsite_bin_chr18, 3
  EM2_H1ESB1_MeDIP_sigleCpG, 4
  EM2_H1ESB2_MeDIP_sigleCpG, 5
  EM_H1ESB1_MeDIP_sigleCpG, 9
  H1ESB1_MRE_sigleCpG, 9
  H1ESB2_MRE_sigleCpG, 10
  three_mre_cpg, 11

* internal
  all_CpGsite_bin_chr18, 3
  classifypvalue, 3
  EM2_H1ESB1_MeDIP_sigleCpG, 4
  EM2_H1ESB2_MeDIP_sigleCpG, 5
  EM_H1ESB1_MeDIP_sigleCpG, 9
  H1ESB1_MRE_sigleCpG, 9
  H1ESB2_MRE_sigleCpG, 10
  three_mre_cpg, 11

* package
  SIMD-package, 2

all_CpGsite_bin_chr18, 3

classifypvalue, 3

EM2_H1ESB1_MeDIP_sigleCpG, 4
EM2_H1ESB2_MeDIP_sigleCpG, 5
EM_H1ESB1_MeDIP_sigleCpG, 9
EMalgorithm, 5
emalgh, 6
emalgh1, 7
EMtest, 7

H1ESB1_MRE_sigleCpG, 9
H1ESB2_MRE_sigleCpG, 10

probBinom, 11
SIMD-package, 2

three_mre_cpg, 11

13