Package ‘ChIPtest’

July 20, 2016

Type Package

Title Nonparametric Methods for Identifying Differential Enrichment Regions with ChIP-Seq Data

Version 1.0

Date 2017-07-07

Author Vicky Qian Wu ; Kyoung-Jae Won ; Hongzhe Li <hongzhe@upenn.edu>

Maintainer Vicky Qian Wu <wuqian7@gmail.com>

Description Nonparametric Tests to identify the differential enrichment region for two conditions or time-course ChIP-seq data. It includes: data preprocessing function, estimation of a small constant used in hypothesis testing, a kernel-based two sample nonparametric test, two assumption-free two sample nonparametric test.

License GPL (>= 2.15.1)

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-07-20 17:59:16

R topics documented:

ChIPtest_1.0-package ................................................. 2
data1 ................................................................. 3
data4 ................................................................. 4
est.c ............................................................... 4
NormTransformation .............................................. 5
TS_kernel ............................................................ 6
TS_twosample ....................................................... 7

Index 9
Nonparametric Methods for Identifying Differential Enrichment Regions with ChIP-seq Data

Description

Nonparametric Tests to identify the differential enrichment region for two conditions or time-course ChIP-seq data. It includes: data preprocessing function, estimation of a small constant used in hypothesis testing, a kernel-based two sample nonparametric test, two assumption-free two sample nonparametric test.

Details

Package: ChIPtest_1.0
Type: Package
Version: 1.0
Date: 2016-07-07
License: GPL (>=2
LazyLoad: yes

Author(s)

Vicky Qian Wu ; Kyoung-Jae Won ; Hongzhe Li <hongzhe@upenn.edu> Maintainer: Vicky Qian Wu <wuqian7@gmail.com>

References


Examples

data(data1)
data(data4)
Data1=NormTransformation(data1)
Data4=NormTransformation(data4)
tao=est.c(Data1, Data4, max1=5, max4=5)
band=54
TS=TS_twosample(Data1, Data4, tao, band, quant=c(0.9,0.9,0.9))
**Description**

In order to identify the genes that show differential histone modification levels between the two conditions, condition A and condition B, ChIPtest consider the upstream 5000 bp region and downstream 2000 bp region around the transcription start site (TSS) for each gene and divide the regions into 280 bins of 25 bps. Since the two ChIP-seq samples are usually sequenced at different depths (total number of reads), the counts were rescaled according to the sequencing depth ratio. In this example, suppose that there are 5 genes and for each gene, there are 280 observations. The input data matrix has 5 rows and 280 columns. Each row represents for one gene, and each column represents for number of short reads covered at one bin after rescaling.

**Usage**

```
data(data1)
```

**Format**

The format is:

```
num [1:5, 1:280] 0 0 0 0 1.43 0 1.43 0 0 1.43 ...
- attr(*, "dimnames")=List of 2
  ..$ : chr [1:5] "1" "2" "3" "4" ...
  ..$ : chr [1:280] "V3" "V4" "V5" "V6" ...
```

**Source**


**References**


**Examples**

```
data(data1)
```
Input data (matrix) for condition B

Description

In order to identify the genes that show differential histone modification levels between the two conditions, condition A and condition B, ChIPtest consider the upstream 5000 bp region and downstream 2000 bp region around the transcription start site (TSS) for each gene and divide the regions into 280 bins of 25 bps. Since the two ChIP-seq samples are usually sequenced at different depths (total number of reads), the counts were rescaled according to the sequencing depth ratio. In this example, suppose that there are 5 genes and for each gene, there are 280 observations. The input data matrix has 5 rows and 280 columns. Each row represents for one gene, and each column represents for number of short reads covered at one bin after rescaling.

Usage

data(data4)

Format

The format is: num [1:5, 1:280] 0 0 0 0 0 0 0 0 0 0 ... - attr(*, "dimnames")=List of 2 ..$ : chr [1:5] "1" "2" "3" "4" ... ..$ : chr [1:280] "V3" "V4" "V5" "V6" ...

Source


References


Examples

data(data4)

calculate the biologically relevant value c in the null hypothesis H0: TS=c, in assumption-free nonparametric test

Description

If there is no INPUT experiment (No control), treat the genes with read counts fewer than 5 as the "null genes". Test statistics were calculated based on the those "null genes" and take the average to obtain the value c, which is used in the null hypothesis H0: TS=c
NormTransformation

Usage

\texttt{est.c(data1, data4, max1 = 5, max4 = 5)}

Arguments

- \texttt{data1} Data Matrix (after VST) for condition A
- \texttt{data4} Data Matrix (after VST) for condition B
- \texttt{max1} Threshold used to decide null genes for condition A. Default as 5
- \texttt{max4} Threshold used to decide null genes for condition B. Default as 5

Details

Data matrix, the default format is N row by M column. Each row represents for one gene, and each column represents for one bin

Value

\texttt{tao} value c in the null hypothesis H0: TS=c

References


Examples

\begin{verbatim}
  data(data1)
data(data4)
Data1=normtransformation(data1)
Data4=normtransformation(data4)
tao=est.c(Data1, Data4, max1=5, max4=5)
\end{verbatim}

\hline
\textbf{NormTransformation} & \textit{Variance-stabilizing transformation (VST) procedure} \\
\hline

Description

Assume observed data approximately follow Poisson Distribution. Apply VST procedure to transform the data as approximate normal with constant variance of 1

Usage

\texttt{NormTransformation(data)}

Arguments

- \texttt{data} Input ChIP-seq data (counts), which approximately follow Poisson Distribution.
Details
Please note the input data can be a single value, a vector or a matrix. If it is a matrix, the default format is N row by M column. Each row represents for one gene, and each column represents for one bin.

Value
After VST transformation, the return data matrix would follow Normal Distribution with a constant variance 1

References

Examples
```r
data(data1)
Data1=NormTransformation(data1)
```

---

**TS_kernel**
*Calculate the Test Statistics for kernel-based nonparametric test.*

Description
Get the difference between two conditions. Apply Kernel smoothing to fit a smooth curve. Estimate variance for each gene and improve the estimation of variance based on all the genes. Derive test statistics and get the rank list of all the genes.

Usage
```r
TS_kernel(data, band, quantile)
```

Arguments
- **data** difference matrix between two conditions
- **band** bandwidth used in kernel smoothing
- **quantile** threshold used in variance estimation

Details
Note 1: Need to chose a bandwidth. Do not recommend to use cross validation (not gene-specific bandwidth) but chose a fixed biological meaningful bandwidth. A fixed bandwidth which can capture the signal profile and smooth out noise would be recommend. The bandwidth used in reference is 20/280.

Note 2: quantile value is based on the distribution of variance estimation of each gene. Recommend to use histogram to double check the distribution. Default 0.9 = 90 %
Value

TS | Kernel based test statistics after WH transformation. Please refer the details in the reference
TS_sign | "+" represent for condition B enriched more than condition A; "-" vice versa
Tmean | Original test statistics, which is calculated as integral of square of kernel estimator

References


Examples

data1=data(data1I
data4=data(dataTI
Data1=NormTransformation(data1)
Data4=NormTransformation(data4)
data=Data4-Data1
band=54
TS=TS_kernel(data, band, quantile=0.9)

TS_twosample Three Nonparametric Test Statistics for two sample ChIP-seq data

Description

It includes three nonparametric test statistics for two sample differential analysis: kernel based nonparametric test, assumption-free nonparametric test with equal variance estimation and unequal variance estimation.

Usage

TS_twosample(data1, data4, tao, band, quant)

Arguments

data1 | data matrix (after VST) for condition A
data4 | data matrix (after VST) for condition B
tao | the biologically relevant value c in the null hypothesis H0: TS=c, in assumption-free nonparametric test
band | bandwidth used in kernel smoothing
quant | threshold used in variance estimation
Details

kernel-based test statistics is the same as "TS_kernel"

Value

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS_kn</td>
<td>kernel based test statistics</td>
</tr>
<tr>
<td>Deq1</td>
<td>assumption-free nonparametric test statistics with equal variance</td>
</tr>
<tr>
<td>Dnum</td>
<td>assumption-free nonparametric test statistics with unequal variance</td>
</tr>
<tr>
<td>sigma1</td>
<td>variance estimation for condition A under equal variance assumption</td>
</tr>
<tr>
<td>sigma4</td>
<td>variance estimation for condition B under unequal variance assumption</td>
</tr>
<tr>
<td>Ts_yvec</td>
<td>Original statistics, which is calculated as integral of square of kernel estimator</td>
</tr>
<tr>
<td>Dsum</td>
<td>Original statistics, which is calculated for nonparametric test without smoothing</td>
</tr>
<tr>
<td>Sev</td>
<td>variance estimation under equal variance assumption</td>
</tr>
<tr>
<td>Suv</td>
<td>variance estimation under unequal variance assumption</td>
</tr>
<tr>
<td>Xg</td>
<td>estimation of standard deviation for kernel-based test statistics</td>
</tr>
</tbody>
</table>

References


Examples

data(data1)
data(data4)
Data1=NormTransformation(data1)
Data4=NormTransformation(data4)
tao=est.c(Data1, Data4, max1=5, max4=5)
band=54
TS=TS_twosample(Data1, Data4, tao, band, quant=c(0.9,0.9,0.9))
Index

*Topic ChIPtest, nonparametric test, ChIP-seq, differential enrichment
  ChIPtest_1.0-package, 2
*Topic datasets
  data1, 3
data4, 4

ChIPtest_1.0 (ChIPtest_1.0-package), 2
ChIPtest_1.0-package, 2
data1, 3
data4, 4

est.c, 4
NormTransformation, 5
TS_kernel, 6
TS_twosample, 7