Package 'exomePeak'

October 9, 2015

Type Package

Title exome-based anlaysis of MeRIP-Seq data: peak calling and differential analysis

Version 1.8.0

Date 2014-09-10

Author Jia Meng <jia.meng@xjtlu.edu.cn>

Maintainer Jia Meng <jia.meng@xjtlu.edu.cn>

Description

The package is developed for the analysis of affinity-based epitranscriptome shortgun sequencing data from MeRIP-seq (maA-seq). It was built on the basis of the exomePeak MATLAB package (Meng, Jia, et al. ``Exome-based analysis for RNA epigenome sequencing data." Bioinformatics 29.12 (2013): 1565-1567.) with new functions for differential analysis of two experimental conditions to unveil the dynamics in post-transcriptional regulation of the RNA methylome. The exomePeak R-package accepts and statistically supports multiple biological replicates, internally removes PCR artifacts and multi-mapping reads, outputs exome-based binding sites (RNA methylation sites) and detects differential post-transcriptional RNA modification sites between two experimental conditions in term of percentage rather the absolute amount. The package is still under active development, and we welcome all biology and computation scientist for all kinds of collaborations and communications. Please feel free to contact Dr. Jia Meng <jia.meng@hotmail.com> if you have any questions.

License GPL-2

Depends Rsamtools, GenomicFeatures (>= 1.14.5), rtracklayer

biocViews Sequencing, HighThroughputSequencing, Methylseq, RNAseq

NeedsCompilation no

R topics documented:

exomePeak-package *exomePeak*

Description

The package is developed for the analysis of affinity-based epitranscriptome shortgun sequencing data from MeRIP-seq (maA-seq). It was built on the basis of the exomePeak MATLAB package (Meng, Jia, et al. "Exome-based analysis for RNA epigenome sequencing data." Bioinformatics 29.12 (2013): 1565-1567.) with new functions for differential analysis of two experimental conditions to unveil the dynamics in post-transcriptional regulation of the RNA methylome. The exomePeak R-package accepts and statistically supports multiple biological replicates, internally removes PCR artifacts and multi-mapping reads, outputs exome-based binding sites (RNA methylation sites) and detects differential post-transcriptional RNA modification sites between two experimental conditions in term of percentage rather the absolute amount. The package is still under active development, and we welcome all biology and computation scientist for all kinds of collaborations and communications. Please feel free to contact Dr. Jia Meng <jia.meng@hotmail.com> if you have any questions.

Details

Package:	exomePeak
Type:	Package
Version:	1.0
Date:	2013-08-02
License:	GPL-2

References

Meng, Jia, Xiaodong Cui, Manjeet K. Rao, Yidong Chen, and Yufei Huang. "Exome-based analysis for RNA epigenome sequencing data." Bioinformatics 29, no. 12 (2013): 1565-1567.

Examples

```
# For usage, please check the main function with:
?exomepeak
```

bltest *bltest*

Description

This is the default test for the differential post-transcriptional RNA modification sites. Differential from all existing tests the compare the absolute amount between two conditions, this test compares whether the percentage of modified molecules are the same.

Usage

```
bltest(untreated_ip, untreated_input,
       treated_ip, treated_input,
       untreated_ip_total, untreated_input_total,
       treated_ip_total, treated_input_total,
       minimal_count_fdr =10)
```

Arguments

`untreated_ip`	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the IP sample under untreated condition
`untreated_input`	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the Input control sample under untreated condition
`treated_ip`	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the IP sample under treated condition
`treated_input`	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the Input control sample under treated condition
`untreated_ip_total`	an integer, total number of reads for the IP sample under untreated condition
`untreated_input_total`	an integer, total number of reads for the Input control sample under untreated condition
`treated_ip_total`	an integer, total number of reads for the IP sample under treated condition
`treated_input_total`	an integer, total number of reads for the Input control sample under treated condition
`minimal_count_fdr`	an integer threshold, only the loci with reads more than this number are subjected for fdr calculation. default: 10

Details

The comparison of 4 Poisson distributions are firstly collapsed into 2 Binomial distributions, and the function further tests whether the two binomial distributions have the same successful rate with a likelihood ratio test. The number of reads at the same locus for the aligned reads are counted by other packages, such as Rsamtools or HTseq-count.

Value

The function returns a list of length 3, which contains the log(p-value), log(fdr) and log(fold change), respectively, from the test.

Author(s)

Lin Zhang, PhD <laurenie.zhang@gmail.com>

References

Reference coming soon!

Examples

```
# input reads count of 3 binding sites
untreated_ip = c(10,20,30)
untreated_input = c(20,20,20)
treated_ip = c(30,10,20)
treated_input  = c(20,20,20)
# sequencing depths
untreated_ip_total = 10^7
untreated_input_total = 10^7
treated_ip_total = 10^7
treated_input_total = 10^7
# get the result
result = bltest(untreated_ip, untreated_input,
treated_ip, treated_input,
untreated_ip_total, untreated_input_total,
treated_ip_total, treated_input_total)
```

---

ctest *ctest*

Description

c-test is used to compare two Poisson means, for peak calling or binding sites identification in exomePeak R-package

Usage

```
ctest(IP, INPUT, TOTAL_IP, TOTAL_INPUT, FOLD = 1, minimal_counts_in_fdr = 10)
```

Arguments

IP	a vector of integers, each elemetnt represents the number of reads from a binding site in the IP sample
INPUT	a vector of integers, each elements represents the number of reads from a binding site in the Input control sample

TOTAL_IP	an integer, which represents the total number of reads in IP sample
TOTAL_INPUT	an integer, which represents the total number of reads in Input control sample
FOLD	a decimal number, which indicates the ration of Possion mean to be tested, default: 1. Use a larger number for detection of highly enriched binding sites.
minimal_counts_in_fdr	an integer threshold, only the loci with reads more than this number are subjected for fdr calculation. default: 10

Details

c-test is used to compare two Poisson means, for peak calling or binding sites identification in exomePeak R-package. The comparison of two Possion distributions is converted into a binomial distribution based test. The number of reads at the same locus for the aligned reads are counted by other packages, such as Rsamtools or HTseq-count.

Value

The function returns a list of length 3, which contains the log(p-value), log(fdr) and log(fold change), respectively.

References

Przyborowski, J. and Wilenski, H. (1940) Homogeneity of results in testing samples from Poisson series: with an application to testing clover seed for dodder. Biometrika, 31, 313-323

Examples

```
result = ctest(c(20,10, 1), c(2,1,20), 100, 200)
```

---

exomepeak *exomepeak*

Description

This is the main function of exomePeak R-package, which supports the processing of affinity-based epitranscriptome sequencing data from MeRIP-Seq (m6A-Seq). The main features of the function includes:

1. Accept and statistically supports multiple biological replicates

2. Remove possible PCR artifacts and mapping ambiguity caused by multi-reads (reads that can be mapped to multiple genomic locations)

3. Peak calling (binding sites detection) and comparison of two experimental conditions (differential analysis)

4. Automatic association of genes and the binding sites; Optionally output the intermediate results in Rdata format

The package features a highly simplied procedure with a single command accomplishing all its functions.

Usage

```
exomepeak(IP_BAM, INPUT_BAM,
          GENOME = NA,
          UCSC_TABLE_NAME = "knownGene",
          GENE_ANNO_GTF = NA,
          TXDB = NA,
          TREATED_IP_BAM = character(0),
          TREATED_INPUT_BAM = character(0),
          OUTPUT_DIR = NA, EXPERIMENT_NAME = "exomePeak_output",
          WINDOW_WIDTH = 200, SLIDING_STEP = 30,
          FRAGMENT_LENGTH = 100, READ_LENGTH = NA,
          MINIMAL_PEAK_LENGTH = FRAGMENT_LENGTH/2,
          PEAK_CUTOFF_PVALUE = NA,
          PEAK_CUTOFF_FDR = 0.05, FOLD_ENRICHMENT = 1,
          CONSISTENT_PEAK_CUTOFF_PVALUE = 0.05,
          CONSISTENT_PEAK_FOLD_ENRICHMENT = 1,
          DIFF_PEAK_METHOD = "rhtest",
          DIFF_PEAK_CUTOFF_PVALUE = NA,
          DIFF_PEAK_CUTOFF_FDR = 0.05,
          DIFF_PEAK_ABS_FOLD_CHANGE = 1,
          DIFF_PEAK_CONSISTENT_CUTOFF_PVALUE = 0.05,
          DIFF_PEAK_CONSISTENT_ABS_FOLD_CHANGE = 1,
          MINIMAL_MAPQ = 30, REMOVE_LOCAL_TAG_ANOMALITIES = TRUE,
          POISSON_MEAN_RATIO = 1, TESTING_MODE = NA,
          SAVE_RESULT_ON_DISK = TRUE)
```

Arguments

IP_BAM	a vector of file names, which specifies a number of IP samples from the untreated condition in bam format
INPUT_BAM	a vector of file names, which specifies a number of Input control samples from the untreated condition in bam format
GENOME	a string,such as "hg19" or "mm9", which specifies the genome assembly used. If a gene annotation file is provided, the exomepeak will call peaks with it; otherwise, exomepeak will download the gene annotation from UCSC using the genome assembly specified here and the gene annotation table specified in "UCSC_TABLE_NAME".
UCSC_TABLE_NAME	a string, which specifies the gene annotation used from UCSC, default: "knownGene". Please use function: supportedUCSCtables() to check available tables. Some tables may not be available for all genomes, and the "refGene" table doesn't work correctly due to multiple occuences of the same transcript on the same chromosome.
GENE_ANNO_GTF	a string, which specifies a gene annotation GTF file if available, default: NA
TXDB	an optional TxDb object for gene annotation information used in the analysis, default: NA. The exomepeak function will first look at TXDB, then GENE_ANNO_GTF,

and then GENOME for gene annnotation information. Please refere to "GenomicFeatures" package for more details about the "TxDb" object.

TREATED_IP_BAM a vector of file names, which specifies a number of IP samples from the treated condition in bam format, default: character(0)

TREATED_INPUT_BAM
a vector of file names, which specifies a number of Input control samples from the treated condition in bam format, default: character(0)

OUTPUT_DIR a string, which specifies the output directory, default: getwd(). By default, exomePeak will output results both 1. as BED/XLS files on disk and 2. returned GRangesList object under the R environment.

EXPERIMENT_NAME
a string, which specifies folder name generated in the output directory that contains all the results, default: "exomePeak_output"

WINDOW_WIDTH an integer, which specifies the window width of the sliding window, default: 200

SLIDING_STEP an integer, which specifies the step of the sliding window, use a smaller number for better resolution, default: 30

FRAGMENT_LENGTH
an integer, which specifies the fragment length in the library preparation, default: 100

READ_LENGTH an integer, which specifies the read length in bam file, default: automatically check the first IP sample

MINIMAL_PEAK_LENGTH
an integer, which specifies the minimal peak length to be reported, default: FRAGMENT_LENGTH/2

PEAK_CUTOFF_PVALUE
a decimal number, which specifies the p-value cut-off in the peak detection algorithm, default: 1e-5

PEAK_CUTOFF_FDR
a decimal number, which specifies the fdr cut-off in the peak detection algorithm. If it is specified, then use "fdr" instead of "p" in peak calling

FOLD_ENRICHMENT
a decimal number, which specifies the minimal fold enrichment in the peak calling process. default: 1

CONSISTENT_PEAK_CUTOFF_PVALUE
used when calling consistent peak. a decimal number, which specifies the p-value cut-off in the peak detection algorithm for each individual sample. All samples must satisfy this cut-off, default: 0.05

CONSISTENT_PEAK_FOLD_ENRICHMENT
used when calling consistent peak. a decimal number, which specifies the fdr cut-off in the peak detection algorithm for each individual sample. All samples must satisfy this cut-off. If it specified, use "fdr" instead of "p"

DIFF_PEAK_METHOD
"bltest" (binomial likelihood ratio test) or "rhtest" (rescaled hypergeometric test), default: "rhtest"

DIFF_PEAK_CUTOFF_PVALUE
: a decimal number, which specifies the p-value cut-off in the comparison of two conditions. If it specified, use "p" instead of "fdr"

DIFF_PEAK_CUTOFF_FDR
: a decimal number, which specifies the fdr cut-off in the comparison of two conditions. default: 0.05

DIFF_PEAK_ABS_FOLD_CHANGE
: a decimal number, which specifies the minimal fold change in the differential analysis. default: 1

DIFF_PEAK_CONSISTENT_CUTOFF_PVALUE
: used when calling consistent differential peak. a decimal number, which specifies the p-value cut-off in the differential peak detection algorithm for each individual sample. All samples must satisfy this cut-off. If it specified, use "p" instead of "fdr".

DIFF_PEAK_CONSISTENT_ABS_FOLD_CHANGE
: used when calling consistent differential peak. a decimal number, which specifies the fdr cut-off in the differential peak detection algorithm for each individual sample. All samples must satisfy this cut-off. default: 0.05

MINIMAL_MAPQ
: the reads used in the analysis, MAPQ "NA" is consider as 255, default: 30

REMOVE_LOCAL_TAG_ANOMALITIES
: a logic variable, which specifies whether remove local tag annomalities, default: TRUE

POISSON_MEAN_RATIO
: a decimal number, which specifies the Poisson mean ratio in ctest, default: 1

TESTING_MODE
: for testing only, an integer used when test whether the package is running correctly, use 100 to get peaks on only the first 100 annotations for a fast test run, default: NA

SAVE_RESULT_ON_DISK
: a logic variable, which indicates whether or not save the result on disk in BED/XLS format as well, default: TRUE. By default, exomePeak will output results both 1. as BED/XLS files on disk and 2. returned GRangesList object under the R environment.

Details

The exomePeak function is an all-in-one command that performs all the core functions of the exomePeak R-package.

For peak calling purpose, it requires the IP and input control samples: An IP sample is the aligned BAM file from the immunoprecipitated sample using RNA modification antibodies such as anti-m6A; The input control sample is the aligned BAM file from the total RNAseq shotgun sequencing.

For differential analysis or comparing two conditions, besides the IP & input samples (from the untreated condition), it also require the IP & input samples from a different condition or the "treated" condition, such as with disease or after subjected to heat shock treatment.

Value

By default, exomePeak will output results both

1. as BED/XLS files on disk (default: "exomePeak_output") under the specified directory (default: current working directory).

2. returned GRangesList object under the R environment.

For the files saved on the disk:

1. If there are only samples from one condition, then detected peaks (RNA methylation sites) and consistent peaks will be reported;

2. If there are samples from two experimental conditions, then detected peaks, significantly differential peaks and consistent differential peaks will be reported in bed and xls formats.

For the returned GRangesList objects:

1. for peak calling when data from one condition is available, the function returns peaks and consistent peaks, and the other information generated in the peak calling process can be accessed with the "mcols" command.

2. for peak calling and differential peaks when data from two condition is available, the function returns peaks, differential peaks on the merged samples (not necessarily consistent on all replicates), and a list of differential peaks consistent for every replicates (recommended list); and the other information generated in differential analysis can be accessed with the "mcols" command.

References

Meng, Jia, Xiaodong Cui, Manjeet K. Rao, Yidong Chen, and Yufei Huang. "Exome-based analysis for RNA epigenome sequencing data." Bioinformatics 29, no. 12 (2013): 1565-1567.

Examples

```
# the exomePeak R-package has two main functions:
# 1. peak detection
# 2. comparison of two conditions
# please feel free to contact jia.meng@hotmail.com for any questions

# specify the parameters
GENE_ANNO_GTF=system.file("extdata", "example.gtf", package="exomePeak")
f1=system.file("extdata", "IP1.bam", package="exomePeak")
f2=system.file("extdata", "IP2.bam", package="exomePeak")
f3=system.file("extdata", "IP3.bam", package="exomePeak")
f4=system.file("extdata", "IP4.bam", package="exomePeak")
IP_BAM=c(f1,f2,f3,f4)
f1=system.file("extdata", "Input1.bam", package="exomePeak")
f2=system.file("extdata", "Input2.bam", package="exomePeak")
f3=system.file("extdata", "Input3.bam", package="exomePeak")
INPUT_BAM=c(f1,f2,f3)
f1=system.file("extdata", "treated_IP1.bam", package="exomePeak")
TREATED_IP_BAM=c(f1)
f1=system.file("extdata", "treated_Input1.bam", package="exomePeak")
TREATED_INPUT_BAM=c(f1)

# peak calling and comparison of two conditions
result = exomepeak(GENE_ANNO_GTF=GENE_ANNO_GTF, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM,
          TREATED_IP_BAM=TREATED_IP_BAM, TREATED_INPUT_BAM=TREATED_INPUT_BAM)
```

```
# or peak calling only, using data from only one condition with the following script
# result = exomepeak(GENE_ANNO_GTF=GENE_ANNO_GTF, IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM)

# alternatively, the gene annotation can be downloaded directly from internet with GENOME (and UCSC_TABLE_NAME).
# this will take a long time with the entire transcriptome of hg19
# result = exomepeak(GENOME="hg19", IP_BAM=IP_BAM, INPUT_BAM=INPUT_BAM)
```

---

rhtest *rhtest*

Description

This is the main test for the differential post-transcriptional RNA modification sites. Differential from all existing tests the compare the absolute amount between two conditions, this test compares whether the percentage of modified molecules are the same.

Usage

```
rhtest(untreated_ip, untreated_input,
       treated_ip, treated_input,
       untreated_ip_total, untreated_input_total,
       treated_ip_total, treated_input_total,
       minimal_count_fdr = 10)
```

Arguments

untreated_ip	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the IP sample under untreated condition
untreated_input	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the Input control sample under untreated condition
treated_ip	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the IP sample under treated condition
treated_input	a vector of integers of n, which is the number of binding sites tested. Each element represents the number of reads fall into a binding site for the Input control sample under treated condition
untreated_ip_total	an integer, total number of reads for the IP sample under untreated condition
untreated_input_total	an integer, total number of reads for the Input control sample under untreated condition
treated_ip_total	an integer, total number of reads for the IP sample under treated condition

`treated_input_total`
: an integer, total number of reads for the Input control sample under treated condition

`minimal_count_fdr`
: an integer threshold, only the loci with reads more than this number are subjected for fdr calculation. default: 10

Details

The `rhtest` function is the main test used in exomePeak for comparing the transcription-independent dynamics in RNA epigenetic regulation between two experimental conditions. The sequencing depth from one condition is rescaled and the reads count from it is rescaled accordingly, so as to apply a hypergeometric test. The number of reads at a specific binding sites for the aligned reads are counted by other packages, such as Rsamtools or HTseq-count.

Value

The function returns a list of length 3, which contains the log(p-value), log(fdr) and log(fold change), respectively, from the test.

References

Meng, Jia, Xiaodong Cui, Manjeet K. Rao, Yidong Chen, and Yufei Huang. "Exome-based analysis for RNA epigenome sequencing data." Bioinformatics 29, no. 12 (2013): 1565-1567.

Examples

```
# input reads count of 3 binding sites
untreated_ip = c(10,20,30)
untreated_input = c(20,20,20)
treated_ip = c(30,10,20)
treated_input  = c(20,20,20)
# sequencing depths
untreated_ip_total = 10^7
untreated_input_total = 10^7
treated_ip_total = 10^7
treated_input_total = 10^7
# get the result
result = rhtest(untreated_ip, untreated_input,
treated_ip, treated_input,
untreated_ip_total, untreated_input_total,
treated_ip_total, treated_input_total)
```

Index