Package 'SCATE'

November 23, 2020

Version 1.0.0

Title SCATE: Single-cell ATAC-seq Signal Extraction and Enhancement

Description SCATE is a software tool for extracting and enhancing the sparse and discrete Single-cell ATAC-seq Signal. Single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) is the state-of-the-art technology for analyzing genome-wide regulatory landscapes in single cells. Single-cell ATAC-seq data are sparse and noisy, and analyzing such data is challenging. Existing computational methods cannot accurately reconstruct activities of individual cis-regulatory elements (CREs) in individual cells or rare cell subpopulations. SCATE was developed to adaptively integrate information from co-activated CREs, similar cells, and publicly available regulome data and substantially increase the accuracy for estimating activities of individual CREs. We demonstrate that SCATE can be used to better reconstruct the regulatory landscape of a heterogeneous sample.

License MIT + file LICENSE

BugReports https://github.com/Winnie09/SCATE/issues

VignetteBuilder knitr

Encoding UTF-8

biocViews ExperimentHub, ExperimentData, Genome, SequencingData, SingleCellData, SNPData

Depends parallel, preprocessCore, splines, splines2, xgboost, SCATEData, Rtsne, mclust

Imports utils, stats, GenomicAlignments, GenomicRanges

Suggests rmarkdown, ggplot2, knitr

RoxygenNote 7.1.1

NeedsCompilation no

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

git_branch RELEASE_3_12

git_last_commit 00a484b

git_last_commit_date 2020-10-27

Date/Publication 2020-11-22

Author Zhicheng Ji [aut],
Weiqiang Zhou [aut],
Wenpin Hou [cre, aut] (<https://orcid.org/0000-0003-0972-2192>),
Hongkai Ji [aut]

Maintainer Wenpin Hou <wp.hou3@gmail.com>

R topics documented:

`cellcluster` *Cell clustering*

Description

Perform Cell Clustering

Usage

```
cellcluster(
  satac,
  type = "reads",
  peakOverlapMethod = "full",
  genome = "hg19",
  clunum = NULL,
  perplexity = 30,
  filtervar = TRUE,
  datapath = NULL
)
```

Arguments

`satac`	If type='reads', satac should be a list of GRanges object of scATAC-seq reads. Each element corresponds to one single cell. The GRanges should be the middle point of the reads with length of 1 base pair. Use 'satacprocess' to preprocess raw reads. If type='peaks', satac should be a list of data frames of scATAC-seq peaks. For each data frame, first column is chromsome name, second column is start site, third column is end site, and fourth column is the number of reads of the peak.
`type`	Character variable of either 'reads' or 'peaks'.
`peakOverlapMethod`	Character variable of either 'full' or 'middle'. Only effective when type = 'peaks'. If peakOverlapMethod='full', then the full range of the peak will be used to find overlap with bins, and all bins overlapping with this peak will be assigned the read counts of this peak. If peakOverlapMethod='middle', only the middle base pair of the peak will be used to find overlap with bins.
`genome`	Character variable of either "hg19" or "mm10".
`clunum`	Numeric variable giving the number of clusters. If NULL the cluster number will be determined automatically

perplexity	Numeric variable specifying perplexity for tSNE. Reduce perplexity when sample size is small.
filtervar	If TRUE, filter out features with low variability.
datapath	Character variable of the path to the customized database (eg myfolder/database.rds). The database can be made using 'makedatabase' function. If not null, 'genome' is ignored.

Details

This function generates averaged signals for CRE clusters and cluster cells.

Value

A list of three components: tsne results, clustering results and aggregated signal for CRE cluster.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
celldata <- lapply(seq_len(50),function(i) {
  pos <- sample(seq_len(1e9),50000)
 GRanges(seqnames=sample(paste0("chr",seq_len(20)),50000,replace=TRUE),IRanges(start=pos,end=pos))
})
names(celldata) <- paste0('cell',seq_len(50))
cellcluster(celldata,type='reads',genome="hg19",filtervar=FALSE,perplexity=1,clunum=3) # reads as input
```

---

extractfeature *Extract features for certain genomic regions*

Description

Extract features for certain genomic regions

Usage

```
extractfeature(res, region, mode = "overlap", folder = NULL)
```

Arguments

res	Matrix outputted by SCATE.
region	Dataframe of genomic region. First column: chromosome name. Second column: start position. Third column: end position.
mode	Either 'overlap' or 'nearest'. If overlap, only bins that overlap with the regions of interest will be shown. If nearest, the nearest bin to each region will be shown. Order will be the same as the input genomic region.
folder	A character value specifying the location where the BED files will be stored. If not NULL, the signal and location of each bin will be saved as BED files in 'folder'. Each cluster has its own BED file, and the name of the BED file is the same as the cluster name. These BED files can be uploaded to UCSC genome browser for visualization.

Details

This function takes as input the recontructed matrix generated by SCATE and a list of genomic regions of interest by the user. It outputs a subset of matrix that overlaps or nearest to the given genomic regions of interest.

Value

If foler is NULL, a subset of the input matrix. Otherwise nothing will be returned and the results will be saved to local folder.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
scateres <- data.frame(combine=seq_len(6))
rownames(scateres) <- paste0('chr1_',c(0:5)*200,'_',199+c(0:5)*200)
extractfeature(scateres,data.frame(seqnames='chr1',start=0,end=201))
```

`makedatabase` *Make customized database*

Description

Make customized database with new bulk DNase-seq data

Usage

```
makedatabase(
  datapath,
  savepath,
  blacklist = NULL,
  bamfile = NULL,
  cre = NULL,
  genome = "hg19",
  genomerange = NULL
)
```

Arguments

`datapath`	path to the data package folder (e.g. myfolder/hg19/). User must first download the data package to use this function. The data package for hg19 and mm10 can be downloaded from http://jilab.biostat.jhsph.edu/projects/scate/hg19.zip or http://jilab.biostat.jhsph.edu/projects/scate/mm10.zip. The compressed file should be unzipped. If users do not want to use existing data compendium (e.g. to build a database in a new species), datapath should be set NULL, and 'genome' will be ignored.
`savepath`	path to save the generated database. e.g. myfolder/database.rds.
`blacklist`	GRanges object that identifies blacklisted regions to be filtered out.
`bamfile`	location of bulk DNase-seq bamfiles.

cre	dataframe of new CRE sites to be added to the database. First column: chromosome name. Second column: start position. Third column: end position.
genome	Character variable of either "hg19" or "mm10". Default is 'hg19'. Ignored when datapath is NULL.
genomerange	Data frame with two columns. First column is the chromosome and second column is the length of the genome. Only useful when datapath is NULL. Example is https://genome.ucsc.edu/goldenpath/help/hg19.chrom.sizes

Details

This function makes a new customized database if users have new bulk DNase-seq data and such information can be contribued to the model building of SCATE.

Value

a new customized database if users have new bulk DNase-seq data and such information can be contribued to the model building of SCATE.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

---

peakcall *Peak calling*

Description

Peak calling function

Usage

```
peakcall(res, flank = 1, fdrcut = 1e-05)
```

Arguments

res	Result matrix returned by SCATE
flank	Numeric variable of the number of flanking bins for each bin. For each bin, an averaged signal of itself and the flanking bins will be calculated and compared to a background distribution.
fdrcut	Numeric variable of FDR cutoff. Bins passing the FDR cutoff will be peaks.

Details

This function performs peak calling for signal generated by SCATE

Value

A list with length equal to the number of clusters. Each element is a data frame with five columns: chromosome name, starting location, ending location, FDR and signal. The data frame is ordered by FDR then by signal.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
gr <- GRanges(seqnames="chr1",IRanges(start=seq_len(100)+1e6,end=seq_len(100)+1e6))
scateout <- SCATE(gr,clunum=156,genome='mm10')[seq_len(1000000),,drop=FALSE]
peakcall(scateout)
```

`satacprocess` *Preprocess scATAC-seq*

Description

Preprocessing scATAC-seq samples

Usage

```
satacprocess(input, type = "bam", libsizefilter = 1000)
```

Arguments

`input`	Either a character vector of locations to bam files (when type is bam) or list of GRanges object of scATAC-seq reads (when type is gr).
`type`	Either 'bam' or 'gr'. 'bam' if the input is locations to bam files. 'gr' if the input is a list of GRanges object.
`libsizefilter`	Numeric variable giving the minimum library size. scATAC-seq samples with library size smaller than this cutoff will be discarded.

Details

This function filters out scATAC-seq with low library size and transforms the reads into middle points of the reads.

Value

GRanges object of list of GRanges object after preprocessing.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
c1 <- GRanges(seqnames="chr1",IRanges(start=seq_len(100)+1e6,end=seq_len(100)+1e6))
c2 <- GRanges(seqnames="chr2",IRanges(start=seq_len(100)+1e6,end=seq_len(100)+1e6))
grl <- list(cell1=c1,cell2=c2)
satacprocess(grl,type='gr',libsizefilter=10)
```

SCATE *Perform SCATE*

Description

Single-cell ATAC-seq signal Extration and Enhancement

Usage

```
SCATE(
  satac,
  type = "reads",
  peakOverlapMethod = "full",
  genome = "hg19",
  cluster = NULL,
  clusterid = NULL,
  clunum = NULL,
  datapath = NULL,
  verbose = TRUE,
  ncores = 1
)
```

Arguments

satac	If type='reads', satac should be a GRanges object or list of GRanges object of scATAC-seq reads. Each element corresponds to one single cell. The GRanges should be the middle point of the reads with length of 1 base pair. Use 'satacprocess' to preprocess raw reads. If type='peaks', satac should be a data frame or list of data frames of scATAC-seq peaks. For each data frame, first column is chromsome name, second column is start site, third column is end site, and fourth column is the number of reads of the peak.
type	Character variable of either 'reads' or 'peaks'.
peakOverlapMethod	Character variable of either 'full' or 'middle'. Only effective when type = 'peaks'. If peakOverlapMethod='full', then the full range of the peak will be used to find overlap with bins, and all bins overlapping with this peak will be assigned the read counts of this peak. If peakOverlapMethod='middle', only the middle base pair of the peak will be used to find overlap with bins.
genome	Character variable of either "hg19" or "mm10". Default is 'hg19'.
cluster	Numeric vector specifying the cluster of cells. Needs to be named and include all cells in satac. If NULL, SCATE will be run on all cells in satac.
clusterid	Numeric number specifying the single cluster to run SCATE. If NULL SCATE will be run on all clusters. Ignored if cluster is NULL. The cluster id must be included in variable 'cluster'.
clunum	Numeric value specifying number of CRE clusters. If NULL, SCATE automatically chooses number of CRE clusters.
datapath	Character variable of the path to the customized database (eg myfolder/database.rds). The database can be made using 'makedatabase' function. If not null, 'genome' is ignored.

verbose	Either TRUE or FALSE. If TRUE, progress will be displayed.
ncores	Numeric variable of number of cores to use. If NULL, the maximum number of cores is used.

Details

This function takes as input the scATAC-seq reads and generates enhanced signals. Users can either perform SCATE on clusters of cells or a single group of cells.

Value

A numeric vector or matrix of values generated by SCATE, depending on the number of clusters. The length of the vector or the number of rows of the matrix is the same as the number of bins in the genome. If a matrix, the column names indicate the cluster id.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
#Reads as input, setting CRE cluster number as 156 to increase speed. Users need to set it to be NULL in real appl
gr <- GRanges(seqnames="chr1",IRanges(start=seq_len(100)+1e6,end=seq_len(100)+1e8))
SCATE(gr,clunum=156,type='reads',genome="mm10")
## Not run:
peak <- data.frame(seqnames="chr1",start=seq_len(100)+1e6,end=seq_len(100)+1e8,count=1)
#Peak as input, peakOverlapMethod=full
SCATE(satac=peak,clunum=156,type='peaks',genome="mm10")
#Peak as input, peakOverlapMethod=middle
SCATE(satac=peak,clunum=156,type='peaks',peakOverlapMethod='middle',genome="mm10")

## End(Not run)
```

SCATEpipeline *SCATE Pipeline*

Description

SCATE pipeline of reading in bam, clustering cell, and performing SCATE

Usage

```
SCATEpipeline(
  bamfile,
  genome = "hg19",
  cellclunum = NULL,
  CREclunum = NULL,
  datapath = NULL,
  ncores = 1,
  perplexity = 30,
  example = FALSE
)
```

Arguments

bamfile	Character vector of bam files to be processed
genome	Character variable of either "hg19" or "mm10".
cellclunum	Numeric variable giving the number of cell clusters when clustering cells. If NULL the cluster number will be determined automatically.
CREclunum	Numeric variable giving the number of CRE clusters when running SCATE. If NULL the cluster number will be determined automatically.
datapath	Character variable of the path to the customized database (eg myfolder/database.rds). The database can be made using 'makedatabase' function. If not null, 'genome' is ignored.
ncores	Numeric variable of number of cores to use. If NULL, the maximum number of cores is used.
perplexity	Numeric variable specifying perplexity of tSNE. Reduce perplexity when sample size is small.
example	An indicator of whether this is running an example or real data. When running a real data, this should be set as FALSE. The default is FALSE.

Details

This function takes as input a list of bam files. It then read in the bam files, cluster cells, and performs SCATE for each cell cluster

Value

A list of three elements. First element is a list generated by cellcluster function, and it contains the cell clustering results. Second element is a matrix generated by SCATE function. Each column is the SCATE result for one cell cluster. Column names indicate the cluster id. Third element is a list of peaks. Each element is the peak list for one cluster. Name of the element indicates the name of the cluster.

Author(s)

Zhicheng Ji, Weiqiang Zhou, Wenpin Hou, Hongkai Ji* <whou10@jhu.edu>

Examples

```
f <- list.files(paste0(system.file(package="SCATEData"),"/extdata/"),full.names = TRUE,pattern='.bam$')
#Users need to set CREclunum to be NULL in real applications.
SCATEpipeline(f[1],genome="hg19",CREclunum=156,perplexity=0.1,example=TRUE)
```

Index