Package ‘EPOM’

January 28, 2018

Type Package

Title EPOM for comparing tissue/cell types based on chromatin states

Version 0.1.0

Author Ruochen Jiang, Wei Vivian Li, Jessica Jingyi Li

Maintainer Ruochen Jiang <ruochenj@gmail.com>

Description It is the R package for the paper:
Li, W.V., Razaee, Z.S. and Li, J.J., 2016, December. Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states. In BMC genomics (Vol. 17, No. 1, p. S10). BioMed Central.
The package includes four steps:
1. Transform bigwig files into a 200 bp matrix.
2. Use ANOVA to select candidate associate regions that have significant difference among cell type groups.
3. Use t test to select associated regions for each cell type group.
4. Calculate EPOM score
You can check EPOM_vignette.pdf for further reference.

License GPL

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

Imports parallel

Suggests knitr, rmarkdown

VignetteBuilder knitr

R topics documented:

anova_select *Use ANOVA to select candidate associated regions*

Description

Step 2 in the EPOM method: use ANOVA to test whether a histone mark has the same group mean signals across different cell and tissue types.

Usage

```
anova_select(cell_type, row_select, matrix_gen, state_name = "enhancer",
  alpha_1 = 1e-10, in_dir = NULL, save = TRUE, out_dir = NULL,
  cores = detectCores() - 1)
```

Arguments

`cell_type`	A character vector indicating the cell type of each sample.
`row_select`	A character vector indicating the row indicies corresponding to a certain state with respect to the original 200 bp matrix.
`matrix_gen`	A matrix which containing all the samples where the signal are correponding to the interval of same length. For example, if we have 127 available samples and set the interval to 200 bp, the dimension of the matrix will be x*(3+127). x is the total length of 200 bp regions. First 3 columns are chrom name (eg. chr1), begin (eg. 201), width (eg. 200) and the other 127 are signals (eg. 0.04) of each sample.
`state_name`	A character string indicating the name for the state. For example, "enhancer1".
`alpha_1`	A number indicating the significance level used in the ANOVA. The default is 1e-10, the larger this number is, the more regions will be selected.
`in_dir`	(Optional) Path of the .RData file to be read in.
`save`	A Boolean indicating whether the output should be saved.
`out_dir`	(Optional) Path of the output of the function.
`cores`	Indicate the number of cores you want to use in your server. The default value is given by function detectCores()-1.

Value

This function returns a list l (lower case "L") including (1) a vector called select_reg which indicates the indices for the regions in the original matrix. (2) a matrix called mat_anova which is the matrix after anova selection.

An RData file called "mat_anova4state.RData" (eg. "mat_anova425.RData" indicates the RData file after processing anova for state 25) includes the object l and l$select_reg corresponds to (1) and l$mat_anova corresponds to (2).

epom — *Calculate the epom scores between every pair of samples.*

Description

This function outputs the epom scores for given arguments. It's package for the paper: Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states by Wei Vivian Li, Zahra S. Razaee and Jingyi Jessica Li (2016).

Usage

```
epom(bw_list, bed_list, bd_header = FALSE, bed_sep = "\t", cell_type,
  histone_mark, state, state_name, alpha_1 = 1e-10, alpha_2 = 0.01,
  m = 13, in_dir = NULL, save = TRUE, out_dir = NULL,
  cores = detectCores() - 1)
```

Arguments

`bw_list` A character vector containing the full paths and file names for all the big wig files. The length and order of bw_list, bed_list and bed_sep should be matched.

`bed_list` A character vector containing the full paths and file names for all the bed files. The length and order of bw_list, bed_list and bed_sep should be matched.

`bd_header` This is used in reading in the bed file. The program uses read.table to read the bed files. This argument corresponds to argument header in read.table. You can use ?read.table for further reference.

`bed_sep` This is used in reading in the bed file. The program uses read.table to read the bed files. This argument corresponds to argument sep in read.table. You can use ?read.table for further reference.

`cell_type` A character vector containing all the cell type information for all the samples. The length and order of bw_list, bed_list and bed_sep should be matched.

`histone_mark` A character string indicating the current histone_mark you are processing

`state` A numeric vector indicating the chromatin states of interest.

`state_name` A character string indicating the name for the state. For example, "enhancer1".

`alpha_1` A number indicating the significance level used in ANOVA. The default is 1e-10, the larger this number is, the more regions will be selected.

`alpha_2` A number indicating the significance level used in the t tests. The default is 0.01.

`m` An integer indicating the significant values required for a region to be selected as the associated region for a cell type.

`in_dir` (Optional) Path of the .RData file to be read in.

`save` A Boolean indicating whether the output should be saved.

`out_dir` (Optional) Path of the output of the function.

`cores` An integer indicating the number of cores to use in parallel computation. The default value is given by function detectCores()-1.

Value

A symmetric matrix containing the epom scores. The rows and columns correspond to the samples, and the matrix entries are the pairwise similarity scores of the samples.

References

Li, Wei Vivian, Razaee, Zahra S. and Li, Jingyi Jessica. "Epigenome overlap measure (EPOM) for comparing tissue/cell types based on chromatin states." BMC genomics 17.1 (2016): S10.

epom_score *Calculate epom_score after the vector chosen by t_test*

Description

Step 4, Calculate epom_score after the vector chosen by t_test.

Usage

```
epom_score(l_t, cell_type, save = TRUE, out_dir = NULL)
```

Arguments

`l_t`	A list returned from step 3 (t_select).
`cell_type`	A character vector containing the cell type information of all the samples.
`save`	A Boolean indicating whether the output should be saved.
`out_dir`	(Optional) Path of the output of the function.

Value

A symmetric matrix containing the epom scores. The rows and columns correspond to the samples, and the matrix entries are the pairwise similarity scores of the samples.

index_num4state *Select the corresponding regions of a specific chromatin state.*

Description

Used in step 1 to select the corresponding regions of a specific chromatin state.

Usage

```
index_num4state(gen_mat, bed, state)
```

Arguments

`gen_mat`	A matrix generated from function matrix_bp_trans.
`bed`	A bed file specifying the corresponding regions for each state.
`state`	A numeric vector indicating the chromatin states of interest.

Value

A vector indicating the rows of the input gen_mat matrix that correspond to a state, whose region information is contained in the input bed file.

`matrix_bp_trans` *Transform BigWig file into a matrix*

Description

Used in step 1 to transform BigWig files into a matrix by averaging the signals within every window with a fixed size (interval).

Usage

```
matrix_bp_trans(bw, interval)
```

Arguments

`bw`	A BigWig file to be transformed.
`interval`	An integer specifying the bandwidth used to average the bigwig signals. We set it to 200 in order to match with the state information provied by bed files.

Value

A matrix with four columns. The four columns includes: chromatin state name (eg. chr1), begin of each region (eg. 201), width of each region (eg. 200) and signal of each region (there is only one column of signal in the output) after transformation.

`t_select` *Use t test to select associated regions of a certian cell type.*

Description

Step 3 in the EPOM method: use t test to identify associated regions of a certain cell type.

Usage

```
t_select(select_reg, mat_anova, cell_type, state_name = "enhancer",
  alpha_2 = 0.01, m = 13, in_dir = NULL, save = TRUE, out_dir = NULL,
  cores = detectCores() - 1)
```

Arguments

`select_reg`	A numeric vector indicating the index of rows corresponding to the orignal transformed matrix after selecing by ANOVA.
`mat_anova`	A matrix containing the signals only on regions selected in ANOVA (rows for regions and columns for samples).
`cell_type`	A character vector indicating the cell type each sample belongs to.
`state_name`	A character string indicating the name for the state. For example, "enhancer1".
`alpha_2`	A number indicating the significance level used in the t tests. The default value is 0.01.

`m` An integer indicating the significant values required for a region to be selected as the associated region for a cell type.

`in_dir` (Optional) Path of the .RData file to be read in.

`save` A Boolean indicating whether the output should be saved.

`out_dir` (Optional) Path of the output of the function.

`cores` Indicate the number of cores you want to use in your server. The default value is given by function detectCores()-1.

Value

A list l (lower case "L") that contains selected enhancer region corresponding to the cell types which have more than one corresponding genes.

An RData file called "enhancer_signal.RData" will be saved corresponding to the result of the t test, each row indicates how t-test results are significant for each cell type.

An RData file called "mat_t.RData" will be saved which includes the list l, where l[[i]] corresponds to the selected region of cell type i. Some of l[[i]] might be NULL if the cell type has only one gene corresponding to it.

Index