# Package 'sharpr2'

May 16, 2018

**Title** Estimating Regulatory Scores and Identifying ATAC-STARR Data

**Version** 1.1.1.0

**Author** Liang He

**Maintainer** Liang He <lianghe@mit.edu>

**Description** An algorithm for identifying high-resolution driver elements for datasets from a high-definition reporter assay library. Xinchen Wang, Liang He, Sarah Goggin, Alham Saadat, Li Wang, Melina Claussnitzer, Manolis Kellis (2017) <doi:10.1101/193136>.

**Depends** R (>= 3.3.0), methods

**Imports** mvtnorm (>= 1.0), Matrix (>= 1.2)

**License** GPL (>= 2)

**Encoding** UTF-8

**LazyData** true

**RoxygenNote** 6.0.1.9000

**Repository** CRAN

**Repository/R-Forge/Project** sharpr2

**Repository/R-Forge/Revision** 17

**Repository/R-Forge/DateTimeStamp** 2018-05-15 23:53:57

**Date/Publication** 2018-05-16 05:06:15 UTC

**NeedsCompilation** no

## R topics documented:

sharpr2-package *Estimating regularoty scores and identifying high resolution driver elements for ATAC-STARR data*

## Description

The package develops an algorithm for identifying high-resolution driver elements for datasets from an ATAC-STARR library.

## Details

| | |
|---|---|
| Package: | sharpr2 |
| Type: | Package |
| Version: | 1.1.1.0000 |
| Date: | 2018-05-12 |
| License: | GPL |

## Author(s)

Liang He

Maintainer: Liang He <lianghe@mit.edu>

## References

High-resolution genome-wide functional dissection of transcriptional regulatory regions in human. Xinchen Wang, Liang He, Sarah Goggin, Alham Saadat, Li Wang, Melina Claussnitzer, Manolis Kellis. bioRxiv 193136; doi: https://doi.org/10.1101/193136

## Examples

```
data(hidra_ex)
re <- sharpr2(hidra_ex[1:2000,], l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=FALSE)
```

call_fdr *call_fdr*

## Description

Calculate FDR-adjusted p-values.

### Usage

```
call_fdr(whole_re, thres_tr = 10, method = 'BH')
```

### Arguments

| | |
|---|---|
| whole_re | A list of the objects obtained from sharpr2 for each chromosome. |
| thres_tr | The threshold for the size of tiled reigons used for calculate FDR-adjusted p-values. The default value is 10. |
| method | The method for calculating FDR-adjusted p-values. See the function 'p.adjust' for more details abount the method. The default is 'BH'. |

### Value

gfdr: a result table (data.frame) containing FDR-adjusted p-values, chromosome, region, the size and the index of the tiled region it is located.

### Examples

```
data(hidra_ex)
whole_re <- sharpr2(hidra_ex, l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=FALSE)
call_fdr(list(whole_re))
```

---

## call_sig_reg *call_sig_reg*

---

### Description

Given an object returned from the sharpr2 function, this function calls significant regions that contain driver elements for a specific tiled region based on a user-defined threshold.

### Usage

```
call_sig_reg(res, nr, threshold = 3.5, win = 10)
```

### Arguments

| | |
|---|---|
| res | An object obtained from the sharpr2 function. |
| nr | An integer indicating the number of tiled region in res for which driver elements will be called. |
| threshold | The cutoff to identify driver elements in the tiled region. The positions with a z-score larger than the threshold will be called. The default is 3.5. |
| win | A window size for removing sporadic significant regions. If a significant consecutive region is small than win, it will be treated as false signals. The default is 10. |

## Value

sig_reg: identified regions containing driver elements.

motif: predicted 20bp core driver elements

## Examples

```
data(hidra_ex)
re  <- sharpr2(hidra_ex[1:2000,], l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=TRUE)
call_sig_reg(re,850, threshold=2.5)
```

---

# call_tile_reg *call_tile_reg*

---

## Description

For a HiDRA dataset on a given chromosome, this function calls tile regions (the regions covered by at least one read).

## Usage

```
call_tile_reg(data)
```

## Arguments

| | |
|---|---|
| data | A data.frame for a HiDRA dataset for one chromosome. The data.frame must contain four columns: 'start', 'end', 'PLASMID', 'RNA', and is sorted by 'start'. |

## Value

tile_reg: A list containing the row ids in the data for each tiled region.

size: The number of reads in each tiled region.

num_r: The total number of tiled regions.

## Examples

```
data(hidra_ex)
tiled <- call_tile_reg(hidra_ex)
```

find_reg *find_reg*

## Description

Given an object from sharpr2 and a position, this function finds the tiled region containing the position.

## Usage

```
find_reg(re,pos)
```

## Arguments

| | |
|---|---|
| re | An object obtained from sharpr2. |
| pos | A position for which the tiled region is searched. |

## Value

ind: the index of the tiled region in the object from sharpr2. If no such tile region is found, NA is returned.

## Examples

```
data(hidra_ex)
re <- sharpr2(hidra_ex[1:2000,], l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=FALSE)
find_reg(re,1000000)
```

hidra_ex *An example dataset including a region of one chromosome from an ATAC-STARR library*

## Description

This is an example dataset containing 10000 fragments with four columns 'start', 'end', 'PLASMID', 'RNA'.

## Usage

```
data(hidra_ex)
```

## Format

The format is a data.frame with the columns: start: the start position of the fragment. end: the end position of the fragment. PLASMID: the count of PLASMID for this fragment. RNA: the count of RNA for this fragment.

## Examples

```
data(hidra_ex)
```

---

`plot.sharpr2` *plot.sharpr2*

---

## Description

Given an object returned from the sharpr2 function, this function plots the estimated scores (with s.e. if available) for a tiled region.

## Usage

```
## S3 method for class 'sharpr2'
plot(x, tr, unc = "CI", loess = FALSE, add = FALSE,
  xlab = "Position", ylab = "Regulatory Score", cicol = 'orange', cimcol = 'grey',
   sreg = TRUE, ...)
```

## Arguments

| | |
|---|---|
| `x` | An object returned from the sharpr2 function. |
| `tr` | An integer indicating which tiled region to be plotted. |
| `unc` | 'MSE' or 'CI', indicating whether to plot sqrt(MSE) or 95%CI for uncertainty. The default is 'CI'. |
| `loess` | An indicator for whether the loess method is used for smoothing in plotting the scores from sharpr2. The standard errors are not plotted when loess is used. |
| `add` | An indicator for whether to add the new plot to the existing one. |
| `xlab` | The label for the x-axis. The default is 'Position'. |
| `ylab` | The label for the y-axis. The default is 'Regulatory Score'. |
| `cicol` | The color for CIs. The default is 'orange'. |
| `cimcol` | The color for filling the regions within CIs. The default is 'grey'. |
| `sreg` | An indicator for whether to highlight the identified driver element regions. The default is TRUE. |
| `...` | Other parameters for plot. |

## Examples

```
data(hidra_ex)
re <- sharpr2(hidra_ex[1:2000,], l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=FALSE)
plot(re,584)
```

sharpr2 *sharpr2*

## Description

For a HiDRA dataset on a given chromosome, this function calls tiled regions (the regions covered by at least one fragment), and calculates regulatory scores for each tiled region. The regulatory scores are based on standardized log(RNA/PLASMID).

## Usage

```
sharpr2(data, l_min = 150, l_max = 600, f_rna = 10, f_dna = 0,
  s_a = 300, verbose = FALSE, auto = TRUE, sig = TRUE, len = FALSE,
  alpha = 0.05, win = 5, mse = FALSE, max_t = 1)
```

## Arguments

| | |
|---|---|
| data | A data.frame containing an ATAC-STARR dataset for one chromosome. The data.frame must contain four columns: 'start', 'end', 'PLASMID', 'RNA'. 'PLASMID' and 'RNA' are the values for DNA and RNA, which should be non-negative real numbers (average value over multiple replicates) or integers (counts). |
| l_min | The fragments with a length smaller than l_min will not be processed. The default is 150. |
| l_max | The fragments with a length larger than l_max will not be processed. The default is 600. |
| f_rna | The fragments with an RNA count smaller than f_rna will not be processed. The default is 10. |
| f_dna | The fragments with an DNA count smaller than f_rna will not be processed. The default is 0. |
| s_a | A variance hyperparameter in the prior for the latent regulatory scores. The default is 1000. |
| verbose | An indicator of whether to show processing information. The default is FALSE. |
| auto | An indicator of whether to automatically estimate the ridge coefficient $\lambda$ from the data for each tiled region using a data-driven way described in the reference. The default is TRUE. If *auto* is TRUE, *s_a* is ignored and a ridge coefficient is estimated for each tiled region separately. If *auto* is FALSE, a global user-defined ridge coefficient (1/s_a) is used. |
| sig | An indicator of whether to identify significant motif regions for the estimated scores. Only valid if auto=TRUE. The default is TRUE. |
| len | An indicator of whether to model log(RNA/PLASMID) of each fragment as the average or the sum of the latent regulatory scores. The default is FALSE, which is the sum. |
| alpha | A regional FWER to call high resolution driver elements (the significant regulatory region). The default is 0.05. |

| | |
|---|---|
| win | A window size for removing sporadic identified significant regions. If a significant consecutive region is small than *win*, it will be treated as false signals. The default is 5. |
| mse | An indicator of whether mean square errors are included in the output results. The default is FALSE. |
| max_t | A value between 0 and 1, indicating the proportion of non-zero eigenvectors used to calculate $\lambda$ when auto=TRUE. The default is 1. |

## Details

The default value of s_a is set to be 300, which is equivalent to a ridge coefficient of 0.0033. This default ridge coefficient value is selected by the median of the estimated $\lambda$ from the first library.

## Value

score: the regulatory scores for each tiled region. This list contains four components: est_a (the regulatory scores at each locus), sd_e (the sqare root of the mean square error), var_nb (the variance of the esitmate at each locus), $\lambda$ (the ridge coefficient).

region: the start and end positions for each tiled region.

n_reg: total number of tiled regions.

n_read: the number of reads in each tiled region.

sig_reg: identified high resolution driver elements based on the cutoff.

motif: predicted 20bp motifs

cutoff: the cutoff used to call high resolution driver elements for the tiled region.

## References

Xinchen Wang, Liang He, Sarah Goggin, Alham Saadat, Li Wang, Melina Claussnitzer, Manolis Kellis. High-resolution genome-wide functional dissection of transcriptional regulatory regions in human.

## Examples

```
data(hidra_ex)
re <- sharpr2(hidra_ex[1:2000,], l_min = 150, l_max = 600, f_dna = 5, f_rna = 0, sig=FALSE)
```

# Index