Package 'TFHAZ'

January 17, 2021

Type Package

Title Transcription Factor High Accumulation Zones

Version 1.12.0

Author Alberto Marchesi, Marco Masseroli

Maintainer Alberto Marchesi <alberto.march91@gmail.com>

Description It finds trascription factor (TF) high accumulation DNA zones, i.e., regions along the genome where there is a high presence of different transcription factors. Starting from a dataset containing the genomic positions of TF binding regions, for each base of the selected chromosome the accumulation of TFs is computed. Three different types of accumulation (TF, region and base accumulation) are available, together with the possibility of considering, in the single base accumulation computing, the TFs present not only in that single base, but also in its neighborhood, within a window of a given width. Two different methods for the search of TF high accumulation DNA zones, called ``binding regions" and ``overlaps", are available. In addition, some functions are provided in order to analyze, visualize and compare results obtained with different input parameters.

License Artistic-2.0

Encoding UTF-8

Imports GenomicRanges, S4Vectors, grDevices, graphics, stats, utils, IRanges, methods

biocViews Software, BiologicalQuestion, Transcription, ChIPSeq, Coverage

LazyData true

Depends R(>= 3.4)

Suggests BiocStyle, knitr, rmarkdown

VignetteBuilder knitr

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

git_branch RELEASE_3_12

git_last_commit 500fb38

git_last_commit_date 2020-10-27

Date/Publication 2021-01-16

R topics documented:

`accumulation` *Creates a vector with accumulation counts of transcription factors for each chromosome base.*

Description

From a dataset with transcription factor (TF) binding regions, this function creates a vector (*accvector*) in which, for each chromosome base, the accumulation of the TFs present in the input dataset is calculated. Three types of accumulation are considered: *TF accumulation*, *region accumulation* and *base accumulation*. TF accumulation: for each base, it is the number of different TFs present in the neighborhood of the considered base. The neighborhood is defined by a window with half-width *w* centered on the considered base. Region accumulation: for each base, it is the number of regions containing TFs in the neighborhood of the considered base. If in the neighborhood of a base there are two input binding regions of the same TF, the accumulation value in that base is equal to 2 (differently from the TF accumulation, whose value in the same case is equal to 1). Base accumulation: for each base, it is the total number of bases belonging to input regions containing TFs in the neighborhood of the considered base. With w=0, a single base approach is applied (no base neighborhood is considered). In this case, if in the input dataset overlapping regions for the same TF and chromosome do not exist, the results of TF, region and base accumulation are equal.

Usage

```
accumulation(data, acctype, chr, w)
```

Arguments

`data`	a GRanges object containing coordinates of TF binding regions and their TF name.
`acctype`	a string with the name of the accumulation type: "TF", "region", "base".
`chr`	a string with the name of the chromosome (e.g., "chr1"). With chr = "all" all the chromosomes in the input GRanges object are considered.
`w`	an integer, half-width of the window that defines the neighborhood of each base.

Value

A list of four elements:

accvector	a Rle (or SimpleRleList if chr = "all") object containing the accumulation for each base of the selected chromosome.
acctype	a string with the accumulation type used.
chr	a string with the chromosome name associated with the accumulation vector.
w	an integer with the half-width of the window used to calculate the accumulation vector.

Examples

```
# loading dataset
data("Ishikawa")
# to calculate TF accumulation for the chromosome 21 for w=0
TF_acc_21_w_0 <- accumulation(Ishikawa,"TF","chr21",0)
```

---

base_dense_w_10 *Contains an output of the dense_zones function.*

Description

base_dense_w_10 is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of *dense_zones* function (with *threshold_step=21* in order to have 14 threshold values) applied to the accumulation vector found with w=10, chr="all", acctype="base". *base_dense_w_10* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# base_dense_w_10 is in the data_man collection of datasets
head(base_dense_w_10)
```

dense_zones *Finds transcription factor dense DNA zones for different accumulation threshold values.*

Description

For each accumulation threshold value defined, this function finds transcription factor (TF) dense DNA zones (regions). Starting from the accumulation vector calculated with the *accumulation* function, each dense zone is formed by contiguous bases with accumulation equal or higher than the threshold. For each defined threshold value, the function finds also the number of dense zones, the number of total bases belonging to the dense zones, the minimum, maximum, mean, median and standard deviation of the dense zone lengths and of the distances between adjacent dense zones.

Usage

```
dense_zones(accumulation, threshold_step, chr = NULL, writeBed = FALSE)
```

Arguments

accumulation	a list of four elements containing: a Rle (or SimpleRleList) object with accumulation values (e.g., obtained with the *accumulation* function), the accumulation type, a chromosome name, and the half-width of the window used for the accumulation count.
threshold_step	an integer, the step used to calculate the threshold values. These values vary from 1 to the maximum accumulation value in the considered accumulation vector (e.g., found with the *accumulation* function). In the case of using accumulation values of base accumulation type and w different from zero (e.g., w=1000), choosing a step quite different from 1 is suggested because the maximum accumulation value is usually high.
chr	optional argument, a string with a cromosome name. It is needed to apply the function only to a single cromosome present in the accumulation input. If chr = "all" (default value) the function operates on all the chromosomes present in the input.
writeBed	When set to TRUE, for each threshold value (and for each chromosome) a ".bed" file with the chromosome and genomic coordinates of the dense zones found is created.

Value

A list of eight elements:

zones	a list with "GRanges" objects with the dense zones found for each chromosome and threshold value considered.
zones_count	a list with a data frame for each chromosome considered, containing the considered threshold values and the number of dense zones obtained with each of the threshold values.
bases_count	a list with a data frame for each chromosome considered, containing the considered threshold values and the total number of bases belonging to the dense zones obtained with each of the threshold values.

lengths	a list with a data frame for each chromosome considered, containing the considered threshold values and min, max, mean, median and standard deviation of the dense zone lengths obtained with each of the considered threshold values.
distances	a list with a data frame for each chromosome considered, containing the considered threshold values and min, max, mean, median and standard deviation of the distances between adjacent dense zones obtained with each of the threshold values.
acctype	a string with the accumulation type used.
chr	a string with the chromosome name associated with the output zones.
w	an integer with half-width of the window used to calculate the accumulation vector.

When writeBed is set to TRUE, for each threshold value (and for each chromosome) a ".bed" file with the chromosome and genomic coordinates of the dense zones found is created.

Examples

```
# loading data
data("data_man")
# TF_acc_w_0 is in the data_man collection of datasets
# to find dense zones, with threshold step equal to 1
TF_dense_w_0 <- dense_zones(TF_acc_w_0, 1)
```

---

high_accumulation_zones

Finds transcription factor high accumulation DNA zones.

Description

This function finds transcription factor high accumulation DNA zones (TFHAZ). Starting from the accumulation vector calculated with the *accumulation* function, two different methods for the search of TF high accumulation DNA zones are available. The *binding regions* method is based on the identification of DNA regions with presence of TF binding (at least one TF) from which those with a high number of different TFs (above the threshold) are selected. This method works only if the accumulation vector is found with *w*=0. The *overlaps* method is the method used also in *dense_zones* function. It uses a single base local approach, identifying DNA bases, that form the dense zones, in which there is high overlap of TFs. For the *binding regions* method the high accumulation zones are the accumulation regions with values higher or equal to the threshold, while in the *overlaps* these zones are defined as sets of contiguous bases with accumulation value higher or equal to the considered threshold. The threshold value is found considering two methods. The *std* method considers all and only the bases of the accumulation vector (*accvector*) with values higher than zero, and the threshold is found with the following formula: *TH = mean(accvector) + 2*std(accvector)*. The *top_perc* method considers the accumulation regions and selects those in the top x percentage, with x chosen by the user through the *perc* argument. The function finds also the number of high accumulation zones, the number of total bases belonging to these zones, the minimum, maximum, mean, median and standard deviation of these zone lengths and of the distances between adjacent high accumulation zones. In the case of *binding regions* method, it is needed to include the *data* input argument, that is the GRanges object used in the *accumulation* function. Furhermore, in the case of single chromosome accumulation vector, the function can plots, for each chromosome base (x axis), the value of accumulation (y axis) calculated with the

accumulation function. On this graph there are also shown the threshold (with a red line) and, on the x axis, the bases belonging to the high accumulation zones (with red boxes). The plot can be saved in a ".png" file.

Usage

```
high_accumulation_zones(accumulation, method = c("overlaps", "binding_regions"),
data, threshold = c("std","top_perc"), perc, writeBed = FALSE, plotZones =
FALSE)
```

Arguments

accumulation	a list of four elements containing: a Rle object (or SimpleRleList) with accumulation values (e.g., obtained with the *accumulation* function), the accumulation type, a chromosome name, and the half-width of the window used for the accumulation count.
method	a string with the name of the method used to find high accumulation zones: "binding_regions" or "overlaps".
data	a GRanges object containing coordinates of TF binding regions and their TF name. It is needed in the case of *binding regions* method.
threshold	a string with the name of the method used to find the threshold value: "std" or "top_perc".
perc	an integer with the percentage value to be used in order to find the threshold with the *top_perc* method.
writeBed	When set to TRUE, for each threshold value a ".bed" file with the chromosome and genomic coordinates of the dense zones found is created.
plotZones	When set to TRUE, and the "accumulation" in input is calculated for a single chromosome, a ".png" file with the plot of the high accumulation zones on the accumulation vector is created.

Value

A list of nine elements:

zones	a GRanges object containing the coordinates of the high accumulation zones.
n_zones	an integer containing the number of high accumulation zones obtained.
n_bases	an integer containing the total number of bases belonging to the high accumulation zones obtained.
lengths	a vector containing the considered threshold value and min, max, mean, median and standard deviation of the high accumulation zone lengths obtained.
distances	a vector containing the considered threshold value and min, max, mean, median and standard deviation of the distances between adjacent high accumulation zones obtained.
TH	a number with the threshold value found.
acctype	a string with the accumulation type used.
chr	a string with the chromosome name associated with the accumulation vector used.
w	an integer with half-width of the window used to calculate the accumulation vector.

Examples

```
# loading dataset
data("Ishikawa")
# TF_acc_w_0 is in the data_man collection of datasets
# to find high accumulation zones
TFHAZ_w_0 <- high_accumulation_zones(TF_acc_w_0, method = "overlaps",
threshold = "std")
```

`Ishikawa` *Contains genomic regions of transcription factors at the ranges side and the name of the transcription factors at the metadata side.*

Description

Ishikawa is a Large GRanges object with 283,009 ranges and 1 metadata column. Each range represents the coordinates of a TF binding region while metadata column indicates the name of the TF.

Usage

```
data("Ishikawa")
```

Format

An object of class "GRanges"

Details

The dataset is obtained from computation of ENCODE ChIP-Seq data of the localization of transcription factor binding sites of the Ishikawa cell line. The data have been processed and extracted with GMQL (GenoMetric Query Language http://www.bioinformatics.deib.polimi.it/ GMQL/).

Value

None, the function is invoked for its side effect.

Examples

```
# loading dataset
data("Ishikawa")
head(Ishikawa)
```

n_zones_PCA *Principal Component Analysis of the number of dense zones obtained with the three methods of accumulation (TF, region, base).*

Description

This function performs the Principal Component Analysis (PCA) of the number of dense zones obtained by varying the threshold on accumulation values obtained with the three methods of accumulation (*TF*, *region*, *base*). Before performing the PCA, the number of dense zone values are scaled with the *scale* R function. This function works only if the number of different threshold values used to find the dense zones with the *dense_zones* function is the same for all the three accumulation types, while the threshold values can be different.

Usage

```
n_zones_PCA(TF_zones, region_zones, base_zones, chr = NULL)
```

Arguments

TF_zones	a list with the results of the *dense_zones* function using the TF accumulation method and varying the thresholds on the considered accumulation values.
region_zones	a list with the results of the *dense_zones* function using the region accumulation method and varying the thresholds on the considered accumulation values.
base_zones	a list with the results of the *dense_zones* function using the base accumulation method and varying the thresholds on the considered accumulation values.
chr	optional argument, needed if the input was found with chr = "all"; a string or a vector containing strings with the name of the chromosome (e.g., "chr1")

Value

A list with a summary containing the standard deviation on each principal component, the proportion of variance explained by each principal component, the cumulative proportion of variance described by each principal component, and the loadings of each principal component. In addition, a plot with the variances of the principal components; a plot with the cumulate variances of the principal components and a plot with the loadings of the three principal components.

Examples

```
# loading data
data("data_man")
# TF_dense_w_10, reg_dense_w_10, and base_dense_w_10 are in the
#data_man collection of datasets
# PCA
n_zones_PCA(TF_dense_w_10, reg_dense_w_10, base_dense_w_10)
```

`plot_accumulation` *Plots the accumulation vector obtained with the accumulation function.*

Description

For each chromosome base (x axis), this function plots the value of accumulation (y axis) calculated with the *accumulation* function. The plot is saved in a ".png" file. If the accumulation input was found with chr = "all", the chromosomes (one or more) to be considered can be chosen.

Usage

```
plot_accumulation(accumulation, chr = NULL)
```

Arguments

`accumulation`	a list of four elements, as the output of the *accumulation* function, containing: a sparse vector with accumulation values, the accumulation type, a chromosome name, and the half-width of the window used for the accumulation count.
`chr`	optional argument, needed if the accumulation input was found with chr = "all"; a string or a vector containing strings with the name of the chromosome(s) (e.g., "chr1" or c("chr1", "chr4")) to be considered.

Value

None, the function is invoked for its side effect.

Examples

```
# loading dataset
data("data_man")
# TF_acc_w_0 is in the data_man collection of datasets
# plot accumulation vector
plot_accumulation(TF_acc_w_0)
```

`plot_n_zones` *Plots the number of dense zones for each threshold value.*

Description

For each accumulation threshold value considered, this function shows the number of dense zones found (e.g., with the *dense_zones* function). It also plots in red color the point of the graph with maximum slope change (maximum second derivative).

Usage

```
plot_n_zones(zones, chr = NULL)
```

Arguments

zones	a list of eight elements, as the output of the *dense_zones* function.
chr	optional argument, needed if the input was found with chr = "all"; a string or a vector containing strings with the name of the chromosome(s) (e.g., "chr1" or c("chr1", "chr4")) to be considered

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_0 is in the data_man collection of datasets
# plot number of zones
plot_n_zones(TF_dense_w_0)
```

---

reg_dense_w_10 *Contains an output of the dense_zones function.*

Description

reg_dense_w_10 is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=10, chr="all", acctype="reg". *reg_dense_w_10* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# reg_dense_w_10 is in the data_man collection of datasets
head(reg_dense_w_10)
```

TF_acc_w_0 *Contains an output of the accumulation function.*

Description

TF_acc_w_0 is a list of 4 elements (*acc_vector*, *acc_type*, *chr*, *w*). It is the output of the accumulation function with acctype="TF", chr="all", w=0. *TF_acc_w_0* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_acc_w_0 is in the data_man collection of datasets
head(TF_acc_w_0)
```

TF_dense_w_0 *Contains an output of the dense_zones function.*

Description

TF_dense_w_0, is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=0, chr="all", acctype="TF". *TF_dense_w_0* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_0 is in the data_man collection of datasets
head(TF_dense_w_0)
```

TF_dense_w_10 *Contains an output of the dense_zones function.*

Description

TF_dense_w_10 is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=10, chr="all", acctype="TF". *TF_dense_w_10* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_10 is in the data_man collection of datasets
head(TF_dense_w_10)
```

TF_dense_w_100 *Contains an output of the dense_zones function.*

Description

TF_dense_w_100 is a list of 7 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=100, chr="all", acctype="TF". *TF_dense_w_100* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_100 is in the data_man collection of datasets
head(TF_dense_w_100)
```

TF_dense_w_1000 *Contains an output of the dense_zones function.*

Description

TF_dense_w_1000 is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=1000, chr="all", acctype="TF". *TF_dense_w_1000* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_1000 is in the data_man collection of datasets
head(TF_dense_w_1000)
```

TF_dense_w_10000 *Contains an output of the dense_zones function.*

Description

TF_dense_w_10000 is a list of 8 elements (*zones*, *zones_count*, *bases_count*, *lengths*, *distances*, *acctype*, *chr*, *w*). It is the output of the *dense_zones* function applied to the accumulation vector found with w=1000, chr="all", acctype="TF". *TF_dense_w_10000* is included in the *data_man* collection.

Usage

```
data("data_man")
```

Format

An object of class list.

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# TF_dense_w_10000 is in the data_man collection of datasets
head(TF_dense_w_10000)
```

---

w_analysis *Shows the number of dense zones obtained with different values of base neighborhood window width.*

Description

This function is used to plot the number of dense zones and the total number of bases belonging to these dense zones present in a set of inputs, obtained (all with accumulation threshold=1) using the *dense_zones* function, for the same accumulation type, same chromosome, and different values of *w* half-width of the window defining the neighborhood of each base. The plot (with x axis logarithmic-scale) is generated only if all input data refer to the same accumulation type, otherwise an error message appears. Beside helping in comparing results obtained with different values of *w*, this function supports finding the best value of *w* to be used in further analysis.

Usage

```
w_analysis(input_list, chr = NULL)
```

Arguments

input_list	a list containing multiple outputs of the *dense_zones* function, obtained for the same accumulation type, same chromosome and different values of *w*.
chr	optional argument, needed if the input was found with chr = "all"; a string or a vector containing strings with the name of the chromosome(s) (e.g., "chr1" or c("chr1", "chr4")) to be considered

Value

None, the function is invoked for its side effect.

Examples

```
# loading data
data("data_man")
# l is a list with dense zone (e.g., TF_dense_w_10, TF_dense_w_100,
# TF_dense_w_1000 and TF_dense_w_10000) objects present in data_man
# collection of datasets
l <- list(TF_dense_w_10, TF_dense_w_100, TF_dense_w_1000,
TF_dense_w_10000)
# plot
w_analysis(l)
```

Index