Package 'similaRpeak'

April 23, 2016

Type Package

Title similaRpeak: Metrics to estimate a level of similarity between two ChIP-Seq profiles

Version 1.2.0

Date 2014-08-08

Description This package calculates metrics which assign a level of similarity between ChIP-Seq profiles.

biocViews BiologicalQuestion, ChIPSeq, Genetics, MultipleComparison, DifferentialExpression

Depends R6 (>= 2.0)

Imports rtracklayer, GenomicAlignments, Rsamtools

Suggests RUnit, BiocGenerics, knitr

License Artistic-2.0 | file LICENSE

URL https://github.com/adeschen/similaRpeak

BugReports https://github.com/adeschen/similaRpeak/issues

VignetteBuilder knitr

NeedsCompilation no

Author Astrid Deschenes [cre, aut], Elsa Bernatchez [aut], Charles Joly Beauparlant [aut], Fabien Claude Lamaze [aut], Rawane Samb [aut], Pascal Belleau [aut], Arnaud Droit [aut]

Maintainer Astrid Louise Deschenes <Astrid-Louise.Deschenes@crchudequebec.ulaval.ca>

R topics documented:

`chr7Profiles` *ChIP-Seq profiles of region chr7:61968807-61969730 related to enhancers H3K27ac and H3K4me1 (for demonstration purpose)*

Description

ChIP-Seq profiles of region chr7:61968807-61969730 of two histone post-transcriptional modifications linked to highly active enhancers H3K27ac (DCC accession: ENCFF000ASG) and H3K4me1 (DCC accession: ENCFF000ARY) from the Encyclopedia of DNA Elements (ENCODE) data (Dunham I et al. 2012).

Usage

```
data(chr7Profiles)
```

Format

A list with 1 entry. The entry is a list of 2 ChIP-Seq profiles, one per active enhancer (H3K27ac and H3K4me1).The 2 ChIP-Seq profiles are of identical length and specific to a genomic region. Each ChiP-Seq profile is a numerical vector containing the profiles values at each position, as reported in reads per million (RPM),

`chr7Profiles` a list containing 1 entry

`chr7Profiles$chr7.61968807.61969730` a list containing 2 ChIP-Seq profiles for the genomic region chr7:6196880-61969730

`demoProfiles$chr7.61968807.61969730$H3K27ac` a numeric vector containing the profiles values related to the enhancer H3K27ac, as reported in reads per million (RPM). The first entry of the vector is for position chr7:61968807 while the last entry is for position chr7:61969730

`demoProfiles$chr7.61968807.61969730$H3K4me1` a numeric vector containing the profiles values related to the enhancer H3K4me1, as reported in reads per million (RPM). The first entry of the vector is for position chr7:61968807 while the last entry is for position chr7:61969730

Source

The Encyclopedia of DNA Elements (ENCODE) data

References

Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74.

See Also

- `MetricFactory` for using a interface to calculate all available metrics separately.
- `similarity` for calculating all available metrics between two ChIP-Seq profiles.

Examples

```
data(chr7Profiles)

## Calculating all metrics for the "chr7.61968807.61969730" region
metrics <- similarity(chr7Profiles$chr7.61968807.61969730$H3K4me1,
                      chr7Profiles$chr7.61968807.61969730$H3K27ac,
                      ratioAreaThreshold=10,
                      ratioMaxMaxThreshold=4,
                      ratioIntersectThreshold=5,
                      ratioNormalizedIntersectThreshold=2,
                      diffPosMaxThresholdMinValue=10,
                      diffPosMaxThresholdMaxDiff=100,
                      diffPosMaxTolerance=0.10)
metrics


## You can refer to the vignette to see more examples using ChIP-Seq profiles
## extracted from the Encyclopedia of DNA Elements (ENCODE) data.
```

---

demoProfiles *Selected ChIP-seq profiles related to enhancers H3K27ac and H3K4me1 (for demonstration purpose)*

Description

ChIP-Seq profiles of two histone post-transcriptional modifications linked to highly active enhancers H3K27ac (DCC accession: ENCFF000ASG) and H3K4me1 (DCC accession: ENCFF000ARY) from the Encyclopedia of DNA Elements (ENCODE) data (Dunham I et al. 2012).

Usage

```
data(demoProfiles)
```

Format

A list with 4 entries. Each entry is a list of 2 ChIP-Seq profiles, one per active enhancer (H3K27ac and H3K4me1).The 2 ChIP-Seq profiles are of identical length and specific to a genomic region. Each ChiP-Seq profile is a numerical vector containing the profiles values at each position, as reported in reads per million (RPM),

`demoProfiles` a list containing all demo ChIP-Seq profiles

`demoProfiles$chr2.70360770.70361098` a list containing 2 ChIP-Seq profiles for the genomic region chr2:70360770-70361098

`demoProfiles$chr2.70360770.70361098$H3K27ac` a numeric vector containing the profiles values related to the enhancer H3K27ac, as reported in reads per million (RPM). The first entry of the vector is for position chr1:70360770 while the last entry is for position chr2:70361098

`demoProfiles$chr2.70360770.70361098$H3K4me1` a numeric vector containing the profiles values related to the enhancer H3K4me1, as reported in reads per million (RPM). The first entry of the vector is for position chr1:70360770 while the last entry is for position chr2:70361098

`demoProfiles$chr3.73159773.73160145$H3K4me1` a list containing 2 ChIP-Seq profiles for the genomic region chr3:73159773-73160145

`demoProfiles$chr3.73159773.73160145$H3K27ac` a numeric vector containing the profiles values related to the enhancer H3K27ac, as reported in reads per million (RPM). The first entry of the vector is for position chr2:73159773 while the last entry is for position chr3:73160145

`demoProfiles$chr3.73159773.73160145$H3K4me1` a numeric vector containing the profiles values related to the enhancer H3K4me1, as reported in reads per million (RPM). The first entry of the vector is for position chr3:73159773 while the last entry is for position chr3:73160145

Source

The Encyclopedia of DNA Elements (ENCODE) data

References

Dunham I, Kundaje A, Aldred SF, et al. *An integrated encyclopedia of DNA elements in the human genome.* Nature. 2012 Sep 6;489(7414):57-74.

See Also

- `MetricFactory` for using a interface to calculate all available metrics separately.
- `similarity` for calculating all available metrics between two ChIP-Seq profiles.

Examples

```
data(demoProfiles)

# Calculate metrics for the "chr3:73159773-73160145" region
metrics <- similarity(demoProfiles$chr3.73159773.73160145$H3K27ac,
                      demoProfiles$chr3.73159773.73160145$H3K4me1)
metrics

## You can refer to the vignette to see more examples using ChIP-Seq profiles
## extracted from the Encyclopedia of DNA Elements (ENCODE) data.
```

MetricFactory-class *MetricFactory object*

Description

An object which is a interface to calculate all available metrics separately.

Details

The `MetricFactory` object is inspired from the factory design pattern. Only one instance of `MetricFactory` object is necessary to calculate all available metrics for different profiles, as long as the thresholds set in the `MetricFactory` instance are appropriate for the calculation. The thresholds are set during the `MetricFactory` object creation and cannot be changed afterwards. If different thresholds are needed, a new `MetricFactory` object, with the new thresholds, must be instantiated.

Value

`MetricFactory$new` returns a `MetricFactory` object which contains the information about the thresholds used to calculate each metric. It can be used, as many times needed, to calculate the specified metrics.

Constructor

`MetricFactory$new( ratioAreaThreshold=1, ratioMaxMaxThreshold=1,`
Create a MetricFactory object.

`ratioAreaThreshold` The minimum denominator accepted to calculate the ratio of the area between both profiles. Default = 1.

`ratioMaxMaxThreshold` The minimum denominator accepted to calculate the ratio of the maximum values between both profiles. Default = 1.

`ratioIntersectThreshold` The minimum denominator accepted to calculate the ratio of the intersection area of both profiles and the total area. Default = 1.

`ratioIntersectThreshold` The minimum denominator accepted to calculate the ratio of the intersection area of both profiles and the total area for normalized profiles. Default = 1.

`diffPosMaxThresholdMinValue` The minimum peak accepted to calculate the metric. Default = 1.

`diffPosMaxThresholdMaxDiff` The maximum distance accepted between 2 peaks positions in one profile to calculate the metric. Default=100.

`diffPosMaxTolerance` The maximum variation accepted on the maximum value to consider a position as a peak position. Default=0.01.

`spearmanCorrSDThreashold` The minimum standard deviation accepted on both profiles to consider to calculate the metric. Default=1e-8.

Author(s)

Astrid Deschenes <Astrid-Louise.Deschenes@crchudequebec.ulaval.ca>

See Also

- `similarity` for calculating all available metrics between two ChIP-Seq profiles.
- `demoProfiles` for more informations about ChIP-Seq profiles present in the demoProfiles data.

Examples

```
## Initialized the factory object
factory = MetricFactory$new(ratioAreaThreshold=100,
                            ratioIntersectThreshold=20,
                            diffPosMaxTolerance=0.04)

## Define 2 ChIP-Seq profiles
profile1=c(1,59,6,24,65,34,15,4,53,22)
profile2=c(15,9,46,44,9,39,27,34,34,4)

## Use the factory object to calculate each metric separatly
ratio_max_max <- factory$createMetric(metricType="RATIO_MAX_MAX",
                                profile1, profile2)
ratio_max_max

diff_pos_max <- factory$createMetric(metric="DIFF_POS_MAX", profile1, profile2)
diff_pos_max

## Example using ChIP-Seq profiles of H3K27ac (DCC accession: ENCFF000ASG)
## and H3K4me1 (DCC accession: ENCFF000ARY) from the Encyclopedia of DNA
## Elements (ENCODE) for the region
data(demoProfiles)

## Visualize ChIP-Seq profiles
plot(demoProfiles$chr3.73159773.73160145$H3K27ac,
        type="l", col="blue", xlab="", ylab="", ylim=c(0, 125),
        main="chr3:73159773-73160145")
par(new=TRUE)
plot(demoProfiles$chr3.73159773.73160145$H3K4me1,
        type="l", col="darkgreen", xlab="Position",
        ylab="Coverage in reads per million (RPM)",  ylim=c(0, 125))
legend("topright", c("H3K27ac","H3K4me1"), cex=1.2,
        col=c("blue","darkgreen"), lty=1)

## Calculate metrics using factory object

ratio_norm_intersect <- factory$createMetric(metricType =
                        "RATIO_NORMALIZED_INTERSECT",
                        profile1=demoProfiles$chr3.73159773.73160145$H3K4me1,
                        profile2=demoProfiles$chr3.73159773.73160145$H3K27ac)
ratio_norm_intersect

ratio_area <- factory$createMetric(metricType="RATIO_AREA",
                        profile1=demoProfiles$chr3.73159773.73160145$H3K4me1,
                        profile2=demoProfiles$chr3.73159773.73160145$H3K27ac)
ratio_area


## You can refer to the vignette to see more examples using ChIP-Seq profiles
## extracted from the Encyclopedia of DNA Elements (ENCODE) data.
```

`similarity` *Calculate metrics which estimate the level of similarity between two ChIP-Seq profiles*

Description

Return a list containing information about both ChIP-Seq profiles and a `list` of all similarity metrics: the ratio of the maximum values, the ratio of the areas, the ratio between the intersection area and the total area (for normalized and non-normalized profiles), the difference between two profiles maximal peaks positions and the Spearman's rho statistic.

Usage

```
similarity(
    profile1,
    profile2,
    ratioAreaThreshold=1,
    ratioMaxMaxThreshold=1,
    ratioIntersectThreshold=1,
    ratioNormalizedIntersectThreshold=1,
    diffPosMaxThresholdMinValue=1,
    diffPosMaxThresholdMaxDiff=100,
    diffPosMaxTolerance=0.01,
    spearmanCorrSDThreashold=1e-8)
```

Arguments

`profile1` Vector containing the RPM values of the first ChIP-Seq profile for each position of the selected region.

`profile2` Vector containing the RPM values of the second ChIP-Seq profile for each position of the selected region.

`ratioAreaThreshold` The minimum denominator accepted to calculate the ratio of the area between both profiles. The value has to be positive. Default = 1.

`ratioMaxMaxThreshold` The minimum denominator accepted to calculate the ratio of the maximal peaks values between both profiles. The value has to be positive. Default = 1.

`ratioIntersectThreshold` The minimum denominator accepted to calculate the ratio of the intersection area of both profiles over the total area. The value has to be positive. Default = 1.

`ratioNormalizedIntersectThreshold` The minimum denominator accepted to calculate the ratio of the intersection area of both normalized profiles over the total area. The value has to be positive. Default = 1.

`diffPosMaxThresholdMinValue`
: The minimum peak accepted to calculate the metric. The value has to be positive. Default = 1.

`diffPosMaxThresholdMaxDiff`
: The maximum distance accepted between 2 peaks positions in one profile to calculate the metric. The value has to be positive. Default=100.

`diffPosMaxTolerance`
: The maximum of variation accepted on the maximum value to consider a position as a peak position. The value can be between 0 and 1. Default=0.01.

`spearmanCorrSDThreashold`
: The minimum standard deviation accepted on both profiles to calculate the metric. Default=1e-8.

Details

`similarity` uses the two vectors passed as arguments to calculate the metrics. When the metric is a ratio, it always verify that the threshold for the denominator is respected. If the threshold is not respected, the metric is assigned the `NA` value.

Value

`similarity` returns a `list` which contains the information about both ChIP-Seq profiles and a `list` of all metrics.

The data structure is a `list` of `list`.

The first level contain the following items:

- nbrPosition: The number of positions included in each profile.
- areaProfile1: The area of the first profile.
- areaProfile2: The area of the second profile.
- maxProfile1: The maximum value in the first profile.
- maxProfile2: The maximum value in the second profile.
- maxPositionProfile1: The list of positions of the maximum value in the first profile.
- maxPositionProfile2: The list of positions of the maximum value in the second profile.
- metrics: A `list` with the following items:
  - RATIO_AREA: The ratio between the areas. The larger value is always divided by the smaller value. `NA` if minimal threshold is not respected.
  - DIFF_POS_MAX: The difference between the maximal peaks positions. The difference is always the first profile value minus the second profile value. `NA` is returned if minimal peak value is not respected. A profile can have more than one position with the maximum value. In that case, the median position is used. A threshold argument can be set to consider all positions within a certain range of the maximum value. A threshold argument can also be set to ensure that the distance between two maximum values is not too wide. When this distance is not respected, it is assumed that more than one peak is present in the profile and `NA` is returned.
  - RATIO_MAX_MAX: The ratio between the maximal peaks values. The first profile is always divided by the second profile. `NA` if minimal threshold is not respected.

- RATIO_INTERSECT: The ratio between the intersection area and the total area. NA if minimal threshold is not respected.
- RATIO_NORMALIZED_INTERSECT: The ratio between the intersection area and the total area of normalized profiles. NA if minimal threshold is not respected.
- SPEARMAN_CORRELATION: The Spearman's rho statistic between profiles. NA if minimal threshold is not respected or when no complete element pair is present between both profiles.

Author(s)

Astrid Deschenes <Astrid-Louise.Deschenes@crchudequebec.ulaval.ca>

Elsa Bernatchez

See Also

- MetricFactory for using a interface to calculate all available metrics separately.
- demoProfiles for more informations about ChIP-Seq profiles present in the demoProfiles data.

Examples

```
## Defining two CHiP-Seq profiles
profile1<-c(3,59,6,24,65,34,15,4,53,22,21,12,11)
profile2<-c(15,9,46,44,9,39,27,34,34,4,3,4,2)

## Example usign default thresholds
similarity(profile1, profile2)

## Example using customised thresholds
similarity(profile1, profile2,
                        ratioAreaThreshold=5,
                        ratioMaxMaxThreshold=5,
                        ratioIntersectThreshold=12,
                        ratioNormalizedIntersectThreshold=2.2,
                        diffPosMaxThresholdMinValue=2,
                        diffPosMaxThresholdMaxDiff=130,
                        diffPosMaxTolerance=0.03,
                        spearmanCorrSDThreashold=1e-3)

## Example using ChIP-Seq profiles of H3K27ac (DCC accession: ENCFF000ASG)
## and H3K4me1 (DCC accession: ENCFF000ARY) from the Encyclopedia of DNA
## Elements (ENCODE) for the region
data(demoProfiles)

## Visualize ChIP-Seq profiles
plot(demoProfiles$chr2.70360770.70361098$H3K27ac,
        type="l", col="blue", xlab="", ylab="", ylim=c(0, 25),
        main="chr2:70360770-70361098")
par(new=TRUE)
plot(demoProfiles$chr2.70360770.70361098$H3K4me1,
        type="l", col="darkgreen", xlab="Position",
```

```
        ylab="Coverage in reads per million (RPM)",  ylim=c(0, 25))
legend("topright", c("H3K27ac","H3K4me1"), cex=1.2,
        col=c("blue","darkgreen"), lty=1)

# Calculate metrics
similarity(demoProfiles$chr2.70360770.70361098$H3K4me1,
            demoProfiles$chr2.70360770.70361098$H3K27ac,
            ratioAreaThreshold=15,
            ratioMaxMaxThreshold=5,
            ratioIntersectThreshold=12,
            ratioNormalizedIntersectThreshold=2.2,
            diffPosMaxThresholdMinValue=2,
            diffPosMaxThresholdMaxDiff=130,
            diffPosMaxTolerance=0.03,
            spearmanCorrSDThreashold=0.1)


## You can refer to the vignette to see more examples using ChIP-Seq profiles
## extracted from the Encyclopedia of DNA Elements (ENCODE) data.
```

`similaRpeak` *similaRpeak: Metrics to estimate a level of similarity between two ChIP-Seq profiles*

Description

similaRpeak is calculating six differents metrics to estimate a level of similarity between two ChIP-Seq profiles

Details

The R function `similarity` calculates six differents metrics:

- RATIO_AREA: The ratio between the areas. The larger value is always divided by the smaller value.
- DIFF_POS_MAX: The difference between the maximal peaks positions. The difference is always a positive value.
- RATIO_MAX_MAX: The ratio between the maximal peaks values. The larger value is always divided by the smaller value.
- RATIO_INTERSECT: The ratio between the intersection area and the total area.
- RATIO_NORMALIZED_INTERSECT: The ratio between the intersection area and the total area of two normalized profiles. The profiles are normalized by divinding them by their average value.
- SPEARMAN_CORRELATION: The Spearman's rho statistic between profiles.

The function `similarity` also reports basic information about each ChIP profile such as the number of positions, the area, the maximum value and the position of the maximum value.

To learn more about **similaRpeak** package see:

https://github.com/adeschen/similaRpeak/wiki

Author(s)

A.L. Deschenes, E. Bertnachez, C. Joly Beauparlant, F.C. Lamaze, R. Samb, P. Belleau and A. Droit

Maintainer: Astrid Louise Deschenes <Astrid-Louise.Deschenes@crchudequebec.ulaval.ca>

See Also

- `MetricFactory` for using a interface to calculate all available metrics separately.
- `similarity` for calculating all available metrics between two ChIP-Seq profiles.

Index