Package 'NarrowPeaks'

September 24, 2012

Version 1.0.1

Date 2012-03-15

Type Package

Title Functional Principal Component Analysis to Narrow Down Transcription Factor Binding Site Candidates

Author Pedro Madrigal <pm@engineering.com>, with contributions from Pawel Krajewski <pkra@igr.poznan.pl>

Description The package processes data in wiggle track format (WIG) commonly produced by several ChIP-seq data analysis tools by applying functional version of principal component analysis (FPCA) over a set of selected candidate enriched regions. This is done in order to shorten the genomic locations accounting for a given proportion of variation among the enrichment-score profiles. It allows the user to discriminate between binding regions in close proximity to each other and to narrow down the length of the putative transcription factor binding sites while preserving the information present in the variability of the dataset and capturing major sources of variation.

Depends R (>= 2.10.0), splines

Maintainer Pedro Madrigal <pm@engineering.com>

Imports GenomicRanges, IRanges, fda, CSAR

Suggests rtracklayer, GenomicRanges, CSAR

License Artistic-2.0

biocViews Visualization, ChIPseq, Transcription, Genetics

LazyLoad yes

LazyData yes

R topics documented:

NarrowPeaks-package *Functional Principal Component Analysis to Narrow Down Transcription Factor Binding Site Candidates*

Description

The package processes data in wiggle track format (WIG) commonly produced by several ChIP-seq data analysis tools by applying functional version of principal component analysis (FPCA) over a set of selected candidate enriched regions. This is done in order to shorten the genomic locations accounting for a given proportion of variation among the enrichment-score profiles. It allows the user to discriminate between binding regions in close proximity to each other and to narrow down the length of the putative transcription factor binding sites while preserving the information present in the variability of the dataset and capturing major sources of variation.

Details

Package:	NarrowPeaks
Type:	Package
Version:	1.0
Date:	2012-03-15
License:	Artistic-2.0
LazyLoad:	yes

Author(s)

Pedro Madrigal, with contributions from Pawel Krajewski <pkra@igr.poznan.pl>

Maintainer: Pedro Madrigal <pm@engineering.com>

References

Madrigal, P. et al. (submitted) NarrowPeaks: an R/Bioconductor package to narrow down transcription factor binding site candidates using functional PCA.

Examples

```
owd <- setwd(tempdir())

##For this example we will use a subset of the AP1 ChIP-seq data (Kaufmann et
##al., 2010)
##The data is obtained after analysis using the CSAR package available in
##Bioconductor
data("NarrowPeaks-dataset")
writeLines(wigfile_test, con="wigfile.wig")

##Write binary files with the WIG signal values for each chromosome
##independently and obtain regions of read-enrichment with score values greater
##than 't', allowing a gap of 'g'. Data correspond to enriched regions found up
##to 105Kb in the Arabidopsis thaliana genome
```

```
wigScores <- wig2CSARScore(wigfilename="wigfile.wig", nbchr = 1,
chrle=c(30427671))
gc(reset=TRUE)
library(CSAR)
candidates <- sigWin(experiment=wigScores$infoscores, t=1.0, g=30)

##Narrow down ChIPSeq enriched regions by functional PCA
shortpeaks <- narrowpeaks(inputReg=candidates,
scoresInfo=wigScores$infoscores, lmin=0, nbf=150, rpenalty=0,
nderiv=0, npcomp=2, pv=80, pmaxscor=3.0, ms=0)

###Export GRanges object with the peaks to annotation tracks in various
##formats. E.g.:
library(GenomicRanges)
names(elementMetadata(shortpeaks$broadPeaks))[3] <- "score"
names(elementMetadata(shortpeaks$narrowPeaks))[2] <- "score"
library(rtracklayer)
export.bedGraph(object=candidates, con="CSAR.bed")
export.bedGraph(object=shortpeaks$broadPeaks, con="broadPeaks.bed")
export.bedGraph(object=shortpeaks$narrowPeaks, con="narrowpeaks.bed")

setwd(owd)
```

---

narrowpeaks *Calculate Narrow Peaks from Enrichment-Score Profiles forming Broad Peaks*

Description

Calculate narrow peaks from enrichment-score profiles forming broad peaks.

Usage

```
narrowpeaks(inputReg, scoresInfo, lmin = 0, nbf = 50, rpenalty= 0,
nderiv= 0, npcomp = 5, pv = 80, pmaxscor = 0.0, ms = 0)
```

Arguments

inputReg	Output of the function sigWin in package CSAR.
scoresInfo	Output infoscores in the function wig2CSARScore, or the function ChIPseqScore after data analysis with package **CSAR**.
lmin	Minimum length of an enriched region from the WIG file to be processed. Integer value.
nbf	Number of order 4 B-spline basis functions that will represent the shape of each candidate site. Integer value.
rpenalty	Smoothing parameter for derivative penalization. Positive numeric value.
nderiv	Order of derivative penalization, if rpenalty>0. Integer value.
npcomp	Number of functional principal components. Integer value greater than or equal to nbf.
pv	Minimum percentage of variation to take into account during the analysis. Numeric value in the range 0-100.

pmaxscor	Cutoff for trimming of scoring function. Numeric value in the range 0-100.
ms	Peaks closer to each other than ms nucleotides are to be merged in the final list. Integer value.

Details

This function produces shortened sites from a list of candidate transcription factor binding sites of arbitrary extension and shape. First, the enrichment signal from each candidate site is represented by a smoothed function constructed using a linear combination of order 4 B-spline basis functions. The data values are fitted using either least squares (if $rpenalty = 0$), or penalized residuals sum of squares (spline smoothing if $rpenalty > 0$).
Then, a functional principal component analysis for npcomp eigenfunctions is performed (Ramsay and Silverman, 2005), giving as a result a set of probe scores (principal component scores) which sum of squares is reported in `elementMetadata(broadPeaks)[,"fpcaScore"]` .The higher the value of fpcaScore, the higher the variance that candidate peak accounts for within the original data. Details on the usage of semi-metrics in functional PCA is described in Ferraty and Vieu, 2006.
After that, we impose the condition that total scoring function for each reported narrow peak must be at least pmaxscor per cent of the maximum value. Max value is calculated from a set of scoring functions using only the eigenfunctions required to achieve pv percent of variance. A new set of scores is computed using trimmed versions of the eigenfunctions (Madrigal et al., submitted), and the root square is stored in `elementMetadata(narrowPeaks)[,"trimmedScore"]`.

Value

A list containing the following elements:

fdaprofiles	A functional data object encapsulating the enrichment profiles (see **fda** package. To plot the data use `plot.fd(fdaprofiles)`).
broadPeaks	Description of the peaks prior to trimming. A GRanges object (see **GenomicRanges** package) with the information: seqnames (chromosome), ranges (start and end of the candidate site), strand (not used), max (maximum signal value for candidate site), average (mean signal value for candidate site), fpcaScore (sum of squares of the first reqcomp principal component scores for candidate site).
narrowPeaks	Description of the peaks after trimming. A GRanges object (see **GenomicRanges** package) with the information: seqnames (chromosome), ranges (start and end after trimming), strand (not used), broadPeak.subpeak, trimmedScore (see details), narrowedDownTo (length reduction relative to the candidate), merged (logical value).
reqcomp	Number of functional principal components used. Integer value.
pvar	Total proportion of variance accounted for by the reqcomp components used. Numeric value in the range 0-100 (always greater than or equal to argument pv).

Author(s)

Pedro Madrigal, <pm@engineering.com>

References

Madrigal, P. et al. (submitted) NarrowPeaks: an R/Bioconductor package to narrow down transcription factor binding site candidates using functional PCA.

Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. New York: Springer.
Ferraty, F. and Vieu, P. (2006) Nonparametric Functional Data Analysis. New York: Springer.

See Also

wig2CSARScore, NarrowPeaks-package

Examples

```
owd <- setwd(tempdir())

##For this example we will use a subset of the AP1 ChIP-seq data (Kaufmann et
##al., 2010)
##The data is obtained after analysis using the CSAR package available in
##Bioconductor
data("NarrowPeaks-dataset")
writeLines(wigfile_test, con="wigfile.wig")

##Write binary files with the WIG signal values for each chromosome
##independently and obtain regions of read-enrichment with score values greater
##than 't', allowing a gap of 'g'. Data correspond to enriched regions found up
##to 105Kb in the Arabidopsis thaliana genome
wigScores <- wig2CSARScore(wigfilename="wigfile.wig", nbchr = 1,
chrle=c(30427671))
gc(reset=TRUE)
library(CSAR)
candidates <- sigWin(experiment=wigScores$infoscores, t=1.0, g=30)

##Narrow down ChIPSeq enriched regions by functional PCA
shortpeaks <- narrowpeaks(inputReg=candidates,
scoresInfo=wigScores$infoscores, lmin=0, nbf=150, rpenalty=0,
nderiv=0, npcomp=2, pv=80, pmaxscor=3.0, ms=0)

###Export GRanges object with the peaks to annotation tracks in various
##formats. E.g.:
library(GenomicRanges)
names(elementMetadata(shortpeaks$broadPeaks))[3] <- "score"
names(elementMetadata(shortpeaks$narrowPeaks))[2] <- "score"
library(rtracklayer)
export.bedGraph(object=candidates, con="CSAR.bed")
export.bedGraph(object=shortpeaks$broadPeaks, con="broadPeaks.bed")
export.bedGraph(object=shortpeaks$narrowPeaks, con="narrowpeaks.bed")

setwd(owd)
```

---

wig2CSARScore *Convert Data from a Wiggle Track (WIG) File to CSAR Binary Format*

Description

Convert data from a wiggle track (WIG) file to CSAR binary format and extract enriched regions.

Usage

```
wig2CSARScore(wigfilename, nbchr, chrle)
```

Arguments

wigfilename	WIG file containing the enrichment-score signal of a transcription factor binding experiment.
nbchr	Number of chromosomes.
chrle	Vector of lengths of the chromosomes (in base pairs).

Details

The Wiggle format (WIG) is described on the UCSC Genome Bioinformatics web site: http://genome.ucsc.edu/FAQ/FAQformat. It allows the display of continuous value data in the genome browser. Although specifically designed for post-processing of WIG files, resulting from the analysis of ChIP-seq experiments (with Bioconductor packages **BayesPeak**, **CSAR**, **PICS**, or other tools such as MACS, F-seq, etc.), **NarrowPeaks** can process other type of sequencing data encoded in WIG format in order to locate regions of high variability in the data.

Value

A list of two elements:

infoscores	A list with the same elements as reported by the function ChIPseqScore in the **CSAR** Bionductor package: chr (Chromosome names), chrL (Chromosome length (bp).), filenames (Name of the files where the score values are stored.), digits (Score values stored on the files need to be divided by 10^digits).

Author(s)

Pedro Madrigal, <pm@engineering.com>

References

Madrigal, P. et al. (submitted) NarrowPeaks: an R/Bioconductor package to narrow down transcription factor binding site candidates using functional PCA.
Muino, J. et al. (2011) ChIP-seq analysis in R (CSAR): An R package for the statistical detection of protein-bound genomic regions. Plant Methods 7:11.

See Also

narrowpeaks, NarrowPeaks-package

Examples

```
owd <- setwd(tempdir())

##For this example we will use a subset of the AP1 ChIP-seq data (Kaufmann et
##al., 2010)
##The data is obtained after analysis using the CSAR package available in
##Bioconductor
data("NarrowPeaks-dataset")
writeLines(wigfile_test, con="wigfile.wig")

##Write binary files with the WIG signal values for each chromosome
##independently and obtain regions of read-enrichment with score values greater
##than 't', allowing a gap of 'g'. Data correspond to enriched regions found up
##to 105Kb in the Arabidopsis thaliana genome
```

```
wigScores <- wig2CSARScore(wigfilename="wigfile.wig", nbchr = 1,
chrle=c(30427671))

setwd(owd)
```

`wigfile_test` *Example Wiggle Track Produced After ChIP-seq Data Analysis*

Description

Sample wiggle track produced after ChIP-seq data analysis. The data represents a small subset of a WIG file storing continuous value scores based on a Poisson test for the chromosome 1 of *Arabidopsis thaliana* (Kaufmann et al., 2010). It contains first 49515 lines of the WIG file for the full experiment.

Format

Wiggle track format (WIG) data in a character vector.

Source

Gene Expression Omnibus GSE20176 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20176). Record from chromatin immunoprecipitation experiments wit AP1-specific antibodies followed by deep-sequencing in order to determine AP1 binding sites on a genome-wide scale in *Arabidopsis thaliana*.

References

Kaufmann et al. (2010) Orchestration of Floral Initiation by APETALA1. Science 328:85-89.

See Also

NarrowPeaks-package

Examples

```
data(NarrowPeaks-dataset)
```

Index