Package 'geneXtendeR'

April 11, 2018

Type Package

Version 1.4.0

Title Optimized Functional Annotation Of ChIP-seq Data

Description geneXtendeR optimizes the functional annotation of ChIP-seq peaks using fast iterative peak-coordinate/GTF alignment algorithms. Since different ChIP-seq peak callers produce different differentially enriched peaks with a large variance in peak length distribution and total peak count, annotating peak lists with their nearest genes can be a noisy process. As such, the goal of geneXtendeR is to robustly link differentially enriched peaks with their respective genes, thereby aiding experimental follow-up and validation in designing primers for a set of prospective gene candidates during qPCR.

Maintainer Bohdan Khomtchouk <khomtchoukmed@gmail.com>

URL https://github.com/Bohdan-Khomtchouk/geneXtendeR

BugReports https://github.com/Bohdan-Khomtchouk/geneXtendeR/issues

Depends GO.db, org.Rn.eg.db, rtracklayer, R (>= 3.3.1)

Imports AnnotationDbi, data.table, dplyr, graphics, networkD3, org.Ag.eg.db, org.Bt.eg.db, org.Ce.eg.db, org.Cf.eg.db, org.Dm.eg.db, org.Dr.eg.db, org.Gg.eg.db, org.Hs.eg.db, org.Mm.eg.db, org.Mmu.eg.db, org.Pt.eg.db, org.Sc.sgd.db, org.Ss.eg.db, org.Xl.eg.db, RColorBrewer, SnowballC, tm, utils, wordcloud

Suggests BiocStyle, knitr, rmarkdown

VignetteBuilder knitr

License GPL (>= 3)

LazyData TRUE

biocViews ChIPSeq, Genetics, Annotation, GenomeAnnotation, DifferentialPeakCalling, Coverage, PeakDetection, ChipOnChip, HistoneModification, DataImport, NaturalLanguageProcessing, Visualization, GO, Software

RoxygenNote 6.0.1

NeedsCompilation yes

Author Bohdan Khomtchouk [aut, cre]

R topics documented:

allPeakLengths . . . 2
annotate . . . 3
barChart . . . 3
cumlinePlot . . . 4
diffGO . . . 5
distinct . . . 5
hotspotPlot . . . 6
linePlot . . . 7
makeNetwork . . . 8
makeWordCloud . . . 9
meanPeakLength . . . 10
meanPeakLengthPlot . . . 10
peakLengthBoxplot . . . 11
peaksInput . . . 12
peaksMerge . . . 12
plotWordFreq . . . 13
rat . . . 14
samplepeaksinput . . . 15

Index **16**

`allPeakLengths` *Produces box-and-whisker plot showing distribution of peak lengths across a peaks input file.*

Description

Makes boxplots of all peak lengths (within a peaks input file) to show how lengths of individual peaks are distributed across the entire peak set.

Usage

```
allPeakLengths(filename)
```

Arguments

`filename`	Name of peaks input file.

Value

Returns a box-and-whisker plot of peak length distribution across a peaks file.

Examples

```
myfile <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
allPeakLengths(myfile)
```

annotate *Annotate peaks file.*

Description

Annotate a user's peaks file (which has been preprocessed with the peaksInput() command) with gene information based on optimally chosen geneXtendeR upstream extension file. This command requires a preprocessed "peaks.txt" file (generated using peaksInput()) to be present in the user's working directory, otherwise the user is prompted to rerun the peaksInput() command in order to regenerate it.

Usage

```
annotate(organism, extension)
```

Arguments

organism	Object name assigned from readGFF() command.
extension	Desired upstream extension.

Value

The gene coordinates are extended by 'extension' at the 5-prime end, and by 500 bp at the 3-prime end. The peaks file is then overlayed on these new gene coordinates, producing a file of peaks annotated with gene ID, gene name, and gene-to-peak genomic distance (in bp). Distance is calculated between 5-prime end of gene and 3-prime end of peak.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
annotate(rat, 2500)
```

barChart *Produces bar charts.*

Description

Makes bar graphs showing the number of genes under peaks at various upstream extension levels.

Usage

```
barChart(organism, start, end, by)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
by	Interval between consecutive extensions.

Value

Creates bar charts.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
barChart(rat, 1000, 3000, 100)
```

---

cumlinePlot *Produces cumulative line plots.*

Description

Makes cumulative differential line plots showing the cumulative sums of the number of genes under peaks at consecutive upstream extension levels.

Usage

```
cumlinePlot(organism, start, end, by)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
by	Interval between consecutive extensions.

Value

Creates cumulative differential line plots.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
cumlinePlot(rat, 1000, 3000, 100)
```

diffGO *Finds differential gene ontologies*

Description

Determines gene ontology terms for each category (biological process (BP), cellular compartment (CC), molecular function (MF)) of genes-under-peaks that are unique between two different upstream extension levels.

Usage

```
diffGO(organism, start, end, GOcategory, GOspecies)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
GOcategory	Either BP, CC, or MF.
GOspecies	Either org.Ag.eg.db (mosquito), org.Bt.eg.db (bovine), org.Ce.eg.db (worm), org.Cf.eg.db (canine), org.Dm.eg.db (fly), org.Dr.eg.db (zebrafish), org.Gg.eg.db (chicken), org.Hs.eg.db (human), org.Mm.eg.db (mouse), org.Mmu.eg.db (rhesus), org.Pt.eg.db (chimpanzee), org.Rn.eg.db (rat), org.Sc.sgd.db (yeast), org.Ss.eg.db (pig), or org.Xl.eg.db (frog).

Value

A data frame of gene symbol, gene ontology ID, and gene ontology term for either a BP, CC, or MF category. This data frame displays the annotations of all unique genes (i.e., genes that are located under peaks between two upstream extension levels) with their respective gene ontology information.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
diffGO(rat, 0, 500, BP, org.Rn.eg.db)
```

distinct *Finds unique genes under peaks.*

Description

Determines what genes directly under peaks are actually unique between two different upstream extension levels.

Usage

```
distinct(organism, start, end)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.

Details

V1-V3 denote the chromosome/start/end positions of the peaks, V4-V6 denote the respective values of the genes, V7 is the gene ID (e.g., Ensembl ID), V8 is the gene name, and V9 is the distance of peak to nearest gene.

Value

A data.table of unique genes located under peaks between two upstream extension levels.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
distinct(rat, 2000, 3000)
```

---

hotspotPlot *Graphs hotspots of statistically significant peak activity.*

Description

Makes line plots showing the ratio of statistically significant peaks to the total number of peaks at each genomic interval (e.g., 0-500 bp upstream of every gene in the genome, 500-1000 bp upstream of every gene in the genome, etc.).

Usage

```
hotspotPlot(totalpeaksfile, significantpeaksfile, organism, start, end, by)
```

Arguments

totalpeaksfile	Filename in user's working directory (or full path to filename) containing all peaks.
significantpeaksfile	Filename in user's working directory (or full path to filename) containing only significant peaks.
organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
by	Interval between consecutive extensions.

Value

Line plot showing the ratio of significant to total peaks at each interval across the genome.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
allpeaks <- system.file("extdata", "totalpeaksfile.txt", package="geneXtendeR")
sigpeaks <- system.file("extdata", "significantpeaksfile.txt", package="geneXtendeR")
hotspotPlot(allpeaks, sigpeaks, rat, 0, 10000, 500)
```

---

linePlot *Produces line plots.*

Description

Makes differential line plots showing the differences in the number of genes under peaks at consecutive upstream extension levels.

Usage

```
linePlot(organism, start, end, by)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
by	Interval between consecutive extensions.

Value

Creates differential line plots.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
linePlot(rat, 1000, 3000, 100)
```

makeNetwork *Makes gene-GO networks*

Description

Creates dynamic and interactive networks of genes linked to their respective gene ontology terms for each category (biological process (BP), cellular compartment (CC), molecular function (MF)) of genes-under-peaks that are unique between two different upstream extension levels.

Usage

```
makeNetwork(organism, start, end, GOcategory, GOspecies)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
GOcategory	Either BP, CC, or MF.
GOspecies	Either org.Ag.eg.db (mosquito), org.Bt.eg.db (bovine), org.Ce.eg.db (worm), org.Cf.eg.db (canine), org.Dm.eg.db (fly), org.Dr.eg.db (zebrafish), org.Gg.eg.db (chicken), org.Hs.eg.db (human), org.Mm.eg.db (mouse), org.Mmu.eg.db (rhesus), org.Pt.eg.db (chimpanzee), org.Rn.eg.db (rat), org.Sc.sgd.db (yeast), org.Ss.eg.db (pig), or org.Xl.eg.db (frog).

Value

An interactive network of gene names linked to their respective gene ontology terms for either a BP, CC, or MF category.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
library(networkD3)
library(dplyr)
library(org.Rn.eg.db)
makeNetwork(rat, 0, 500, BP, org.Rn.eg.db)
```

makeWordCloud *Makes word cloud from gene ontology terms*

Description

Creates word cloud from gene ontology terms derived from either biological process (BP), cellular compartment (CC), or molecular function (MF) of genes-under-peaks that are unique between two different upstream extension levels.

Usage

```
makeWordCloud(organism, start, end, GOcategory, GOspecies)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
GOcategory	Either BP, CC, or MF.
GOspecies	Either org.Ag.eg.db (mosquito), org.Bt.eg.db (bovine), org.Ce.eg.db (worm), org.Cf.eg.db (canine), org.Dm.eg.db (fly), org.Dr.eg.db (zebrafish), org.Gg.eg.db (chicken), org.Hs.eg.db (human), org.Mm.eg.db (mouse), org.Mmu.eg.db (rhesus), org.Pt.eg.db (chimpanzee), org.Rn.eg.db (rat), org.Sc.sgd.db (yeast), org.Ss.eg.db (pig), or org.Xl.eg.db (frog).

Value

A word cloud comprised of words gathered from gene ontology terms of either a BP, CC, or MF category.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
makeWordCloud(rat, 0, 500, BP, org.Rn.eg.db)
```

`meanPeakLength` *Calculates mean (average) peak length for any genomic region.*

Description

Determines the average peak length of all peaks found within some genomic interval (e.g., 0-500 bp upstream of nearest gene for all genes throughout the genome).

Usage

```
meanPeakLength(organism, start, end)
```

Arguments

`organism`	Object name assigned from readGFF() command.
`start`	Lower bound of upstream extension.
`end`	Upper bound of upstream extension.

Value

A vector composed of a single number representing the average peak length found within a genomic interval.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
sigpeaks <- system.file("extdata", "significantpeaksfile.txt", package="geneXtendeR")
peaksInput(sigpeaks)
meanPeakLength(rat, 0, 500)
```

`meanPeakLengthPlot` *Produces line plots of mean (average) peak length within any genomic interval.*

Description

Makes line plots of mean peak lengths to show the average length of individual peaks within any genomic interval (e.g., 0-500 bp upstream of nearest gene for all genes throughout the genome).

Usage

```
meanPeakLengthPlot(organism, start, end, by)
```

Arguments

`organism`	Object name assigned from readGFF() command.
`start`	Lower bound of upstream extension.
`end`	Upper bound of upstream extension.
`by`	Interval between consecutive extensions.

Value

Creates mean peak length line plots.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
allpeaks <- system.file("extdata", "totalpeaksfile.txt", package="geneXtendeR")
peaksInput(allpeaks)
meanPeakLengthPlot(rat, 0, 10000, 500)
```

---

peakLengthBoxplot *Produces box-and-whisker plot of peak lengths within any genomic interval.*

Description

Makes boxplots of peak lengths to show how lengths of individual peaks are distributed within any genomic interval (e.g., 0-500 bp upstream of nearest gene for all genes throughout the genome).

Usage

```
peakLengthBoxplot(organism, start, end)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.

Value

Creates boxplots showing how lengths of peaks are distributed within any given genomic interval. Also, creates character vector composed of individual peak lengths.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
allpeaks <- system.file("extdata", "totalpeaksfile.txt", package="geneXtendeR")
peaksInput(allpeaks)
peakLengthBoxplot(rat, 0, 500)
```

peaksInput *Preprocesses a peaks input file.*

Description

Takes your tab-delimited 3-column (chromosome number, peak start, and peak end) input file (see ?samplepeaksinput) consisting of peaks called from a peak caller (e.g., MACS2 or SICER) and sorts the file by chromosome and start position, thereby creating a preprocessed file for downstream geneXtendeR analysis. This file (called "peaks.txt") is a preprocessed file of the original input and is deposited in the user's working directory and used for the remainder of the analysis. In this "peaks.txt" file, peaks are sorted by chromosome number and start position, where the X chromosome is designated by the integer 100, the Y chromosome by the integer 200, and the mitochondrial chromosome by the integer 300.

Usage

```
peaksInput(filename)
```

Arguments

filename	Name of file containing peaks that have been generated from a peak caller (e.g., MACS2, SICER). See ?samplepeaksinput for an example of such an input file.

Value

Returns a formatted file (called "peaks.txt") that has been preprocessed in preparation for usage with barChart(), linePlot(), distinct(), and other downstream commands and deposited in the user's working directory.

Examples

```
?samplepeaksinput  #Documentation of some exemplary sample input
data(samplepeaksinput)
head(samplepeaksinput)
tail(samplepeaksinput)
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
```

peaksMerge *Transform peaks into merged peaks.*

Description

Takes your tab-delimited 3-column (chromosome number, peak start, and peak end) input file (see ?samplepeaksinput) consisting of peaks called from a peak caller (e.g., MACS2 or SICER) and transforms this file into a file of merged peaks. This file (called "peaks.txt") is a preprocessed file of the original input transformed into merged peaks, and it is deposited in the user's working directory and used for the remainder of the analysis. In this "peaks.txt" file, peaks are sorted by chromosome number and start position, where the X chromosome is designated by the integer 100, the Y chromosome by the integer 200, and the mitochondrial chromosome by the integer 300.

Usage

```
peaksMerge(filename, mergeby)
```

Arguments

filename	Name of file containing peaks that have been generated from a peak caller (e.g., MACS2, SICER). See ?samplepeaksinput for an example of such an input file.
mergeby	Integer indicating how close two adjacent sorted peaks need to be in order to be merged into one peak.

Value

Returns a formatted file (called "peaks.txt"), deposited in the user's working directory, which has been preprocessed to transform individual peaks into merged peaks in preparation for usage with barChart(), linePlot(), distinct(), and other downstream commands.

Examples

```
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksMerge(fpath, 500)
```

---

plotWordFreq *Plots word frequencies found within gene ontology terms*

Description

Creates barplots of word frequencies from gene ontology terms derived from either biological process (BP), cellular compartment (CC), or molecular function (MF) of genes-under-peaks that are unique between two different upstream extension levels.

Usage

```
plotWordFreq(organism, start, end, GOcategory, GOspecies, word_count)
```

Arguments

organism	Object name assigned from readGFF() command.
start	Lower bound of upstream extension.
end	Upper bound of upstream extension.
GOcategory	Either BP, CC, or MF.
GOspecies	Either org.Ag.eg.db (mosquito), org.Bt.eg.db (bovine), org.Ce.eg.db (worm), org.Cf.eg.db (canine), org.Dm.eg.db (fly), org.Dr.eg.db (zebrafish), org.Gg.eg.db (chicken), org.Hs.eg.db (human), org.Mm.eg.db (mouse), org.Mmu.eg.db (rhesus), org.Pt.eg.db (chimpanzee), org.Rn.eg.db (rat), org.Sc.sgd.db (yeast), org.Ss.eg.db (pig), or org.Xl.eg.db (frog).
word_count	Number of top words to display

Value

A barplot comprised of words sorted by frequency of occurrence from gene ontology terms of either a BP, CC, or MF category.

Examples

```
rat <- readGFF("ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norvegicus.Rnor_6.0.84.gt
fpath <- system.file("extdata", "somepeaksfile.txt", package="geneXtendeR")
peaksInput(fpath)
library(tm)
library(SnowballC)
library(wordcloud)
library(RColorBrewer)
plotWordFreq(rat, 0, 500, BP, org.Rn.eg.db, 10)
```

---

rat *Gene transfer format (GTF) file for rat (Rattus_norvegicus.Rnor_6.0.84)*

Description

A dataset downloaded from Ensembl that contains the entries of a GTF file for Rattus norvegicus.

Usage

```
data(rat)
```

Format

A data frame with 748514 rows and 28 variables corresponding to the entries of a GTF file. Column names are standardized and can be found here: http://www.ensembl.org/info/website/upload/gff.html.

Value

Demonstrates a rat GTF file downloaded from: ftp://ftp.ensembl.org/pub/release-84/gtf/rattus_norvegicus/Rattus_norveg

Examples

```
head(rat)
tail(rat)
```

samplepeaksinput *Sample peaks list to be used as input to geneXtendeR*

Description

A dataset containing the chromosome number, start and stop positions of ChIP-seq peaks along the Rattus norvegicus genome (rn6 assembly). A dataset like this may be used as input to the peaksInput() command, which will sort the dataset by chromosome number and start position.

Usage

```
data(samplepeaksinput)
```

Format

A data frame with 25089 rows and 3 variables:

chr Chromosome number

start Peak start position [in units of base pairs]

end Peak end position [in units of base pairs]

Value

Demonstrates a sample peaks file used as input.

Examples

```
head(samplepeaksinput)
tail(samplepeaksinput)
```

Index

∗Topic **datasets**
rat, 14
samplepeaksinput, 15

allPeakLengths, 2
annotate, 3

barChart, 3

cumlinePlot, 4

diffGO, 5
distinct, 5

hotspotPlot, 6

linePlot, 7

makeNetwork, 8
makeWordCloud, 9
meanPeakLength, 10
meanPeakLengthPlot, 10

peakLengthBoxplot, 11
peaksInput, 12
peaksMerge, 12
plotWordFreq, 13

rat, 14

samplepeaksinput, 15