Package 'AIRs'

February 6, 2018

Type Package

Title Analyzer of Integrated Regions

Version 0.99.0

Date 2018-02-06

Author@R c(person(``Min-Jeong'', ``Baek'', email = ``mjbaek16@korea.ac.kr'', role = c(``aut'', ``cre'')), person(``In-Geol'', ``Choi'', email = ``igchoi@korea.ac.kr'', role = c(``aut'')))

Author Min-Jeong Baek [aut, cre], In-Geol Choi [aut]

Maintainer Computational & Synthetic Biology Lab <igchoi@korea.ac.kr>

Description This package was developed for analysis of regions where viral vectors are integrated. Find the location of integrated regions and analyze whether it is associated with important genomic factors.
Finally, user can conduct random analysis based on the results of the previous analysis.

Depends R (>= 3.2.5)

Imports data.table (>= 1.10.4), ggbio (>= 1.18.5), ggplot2 (>= 2.2.1), GenomeInfoDb (>= 1.6.3), GenomicRanges (>= 1.22.4), grDevices (>= 3.2.5), graphics (>= 3.2.5), IRanges (>= 2.4.8), seqinr (>= 3.3), stats (>= 3.2.5), stringr (>= 1.2.0), S4Vectors (>= 0.8.11), utils (>= 3.2.5)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

R topics documented:

`canine` *Refseq & UCSC chromosome id list - Canine*

Description

Data is made from target FASTA file from NCBI Refseq and UCSC genome browser database.

Usage

```
data(canine)
```

Format

An object of class `data.frame` with 40 rows and 2 columns.

Examples

```
data(canine)
```

`chicken` *Refseq & UCSC chromosome id list - Chicken*

Description

Data is made from target FASTA file from NCBI Refseq and UCSC genome browser database.

Usage

```
data(chicken)
```

Format

An object of class `data.frame` with 8064 rows and 2 columns.

Examples

```
data(chicken)
```

`distribution` *Functions for distribution analysis*

Description

This function is for distribution analysis. Outputs are two kinds of histogram, drawing frequency and density, and they are saved in outputpath. Plus, Each of data is saved in a data frame format variable.

Usage

```
distribution(rawHitTable, annoHitTable, annotTableType, isHuman = FALSE,
  featureTable, interval = 2, outputpath = getwd())
```

Arguments

`rawHitTable`	Data frame of BLAST result (from makeHitTable_output$rawHits).
`annoHitTable`	Data frame of annotated hits (from makeHitTable_output$annoHits).
`annotTableType`	Choose one of genetical features such as gene, cpg, repeat, tss.
`isHuman`	If target genome is human, enter TRUE. Default is FALSE.
`featureTable`	Data frame of annotation databases.
`interval`	Interval of distribution graph. User can choose 2kb or 5kb. Default is 2kb.
`outputpath`	Full path of output file located (Except for file name). Default is working directory.

Format

Output format is a list :

overlaps A data table which consists of hits which have identities more than 95.

histdata A data table annotated form.

See Also

`findHits`

`makeHitTable`

Examples

```
distribution(hitset$rawHits, hitset$annotHits, annotTableType = cpg, featureTable = cpgdb,
             interval = 2, outputpath = ~/test)
```

drawingIdeo *Function for drawing an ideogram plot*

Description

This function makes an ideogram of NCBI Refseq chromosomes. An Ideogram image is saved in the output folder.

Usage

```
drawingIdeo(regions, genes, rawHitTable, blastdb = "ncbi", chridTable,
  outputpath = getwd())
```

Arguments

regions	Data frame from NCBI Refseq annotation file (from extractFeatures_output$regions).
genes	Data frame from annotation database file. (from extractFeatures_output$genes).
rawHitTable	Data frame of BLAST result (from makeHitTable_output$rawHits).
blastdb	Used sequence data for running BLASTm. User can choose ncbi and ucsc. Default is ncbi.
chridTable	A Data frame from package's chrid datasets.
outputpath	Full path of output file located (Except for file name). Default is working directory.

See Also

findHits

makeHitTable

extractFeatures

Examples

```
drawingIdeo(anno$region, anno$genes, hitset$rawHits, blastdb = ncbi, chicken, outputpath = ~/test)
```

`extractFeatures` *Function for making various type data table from NCBI Refseq annotation file.*

Description

This function makes data table include NCBI annotation features for searching hits' information. Before use this function, user should generate a data frame of gff file using by readGFF. Outputs are 4 data frames about region, genes, transcripts and extra features. Especially, transcript data of output is used for transcription start site analysis of not human and mouse.

Usage

```
extractFeatures(gff)
```

Arguments

`gff`	A Data frame of GFF file.

Format

Output format is a list :

region A data table which consists of hits which have 'Region' feature.

genes A data table which consists of hits which have 'Gene' feature.

transcripts A data table which consists of hits which have 'Transcript' or 'RNA' features.

etc A data table which consists of hits which are not included previous sets.

See Also

`readGFF`

`drawingIdeo`

Examples

```
extractFeatures(gff)
```

FASTAinfo *Function for profiling FASTA file*

Description

It is a function that displays sequence information in a fasta file.

Usage

```
FASTAinfo(inputpath)
```

Arguments

inputpath	Full path of input file (FASTA format).

Format

Output is a data frame with 4 columns :

num The number of sequence in this file.

avr The average length of sequences in this file.

min The minimum value of sequence length.

max The maximum value of sequence length.

Examples

```
FASTAinfo(inputpath = ~/test/chicken.fna)
```

findHits *Function for finding location of integration sites*

Description

This function allows you to run CD-HIT-EST to reduce redundants and nucleotide BLAST for searching integrated sites. For using this function, CD-HIT-EST and BLASTn is already installed.

Usage

```
findHits(inputfile, outputpath = getwd(), cdhitpath, blastnpath, blastdbpath,
  btask = "megablast", thread = 1)
```

Arguments

inputfile	Full path of sequence file (FASTA format).
outputpath	Full path of output file located (Except for file name). Default is working directory.
cdhitpath	Full path of CD-HIT-EST installed.
blastnpath	Full path of BLASTn installed.
blastdbpath	Full path of blastdb saved.
btask	Choose a task between blastn and megablast. Default is megablast.
thread	The number of core in your server for running BLASTn. Default is 1.

See Also

makeHitTable

distribution

random

drawingIdeo

Examples

```
findHits(inputfile = ~/test/chicken.fna, outputpath = ~/test,
         cdhitpath = /csbl_local/tools/ngs/cd-hit-v4.6.6-2016-0711/cd-hit-est,
         blastnpath = /csbl_local/tools/ngs/ncbi-blast-2.5.0+/bin/blastn,
        blastdbpath = /home/shiny/irahome/db/chicken.fna, btask = megablast, thread = 4)
```

---

human *Refseq & UCSC chromosome id list - Human*

Description

Data is made from target FASTA file from NCBI Refseq and UCSC genome browser database.

Usage

```
data(human)
```

Format

An object of class `data.frame` with 328 rows and 2 columns.

Examples

```
data(human)
```

makeHitTable *Function for making hit table about integrated regions into genomic factors.*

Description

This function selects hits integrated into genomic factors and these are from BLAST result file. User can control minimum value of identities.

Usage

```
makeHitTable(inputdata, inputformat = "dataframe", ident_value = 95,
  chridTable, annotTableType, featureTable)
```

Arguments

inputdata	Full path of BLAST file (TBL format) or Data from findHits function (Dataframe).
inputformat	Format of input data. User can choose one The user selects one of tbl and dataframe. Default is dataframe.
ident_value	Number of identities to filter out blast hits. Default is 95.
chridTable	A Data frame from package's chrid datasets.
annotTableType	Choose one of genetical features such as gene, cpg, repeat, tss. If you are running tss annotation, you can only use it for non-human sequence analysis.
featureTable	A Data frame from annotation databases.

Format

Output format is a list :

rawHits A data table which consists of hits which have identities more than specific value.

annoHits A data table annotated form.

multiHits A data table showed multihits.

See Also

extractFeatures

findHits

GenomicRanges

Examples

```
makeHitTable(inputdata = ~/test/blastn_result/chicken.tbl, inputformat = tbl, chridTable = chicken,
             annotTableType = cpg, featureTable = cpgdb)
```

monkey *Refseq & UCSC chromosome id list - Monkey*

Description

Data is made from target FASTA file from NCBI Refseq and UCSC genome browser database.

Usage

```
data(monkey)
```

Format

An object of class `data.frame` with 572 rows and 2 columns.

Examples

```
data(monkey)
```

mouse *Refseq & UCSC chromosome id list - Mouse*

Description

Data is made from target FASTA file from NCBI Refseq and UCSC genome browser database.

Usage

```
data(mouse)
```

Format

An object of class `data.frame` with 44 rows and 2 columns.

Examples

```
data(mouse)
```

random *Function for random analysis*

Description

This function generates random samples and do chi-square test to see similarity. By this function, you can make random distribution graph and chi-square statistic values.

Usage

```
random(overlapTable, rawHitTable, annoHitTable, histdata, featureTable,
  interval = 2, samplenum = 1e+05, outputpath = getwd())
```

Arguments

overlapTable	Data frame of findOverlaps output (from distribution_output$overlaps).
rawHitTable	Data frame of BLAST result (from makeHitTable_output$rawHits).
annoHitTable	Data frame of annotated hits (from makeHitTable_output$annoHits).
histdata	Data table of distribution (from distribution_output$histdata).
featureTable	Data frame of annotation databases.
interval	Interval of distribution graph. User can choose 2kb or 5kb. Default is 2kb.
samplenum	The number of samples in random set. Default value is 100000.
outputpath	Full path of output file located (Except for file name). Default is working directory.

Format

Output format is a list :

random A random set.

chitable A data table that shows chitest result.

chiresult Summary of chi-squre test(by chisq.test()).

See Also

findHits

makeHitTable

distribution

chisq.test

Examples

```
random(overlapTable = distr$overlaps, rawHitTable = hitset$rawHits, annoHitTable = hitset$annotHits,
     histdata = distr$histdata, featureTable = cpgdb, samplenum = 100000, outputpath = ~/test)
```

readCpGdb — *Function for converting annotation file from UCSC to specific data frame*

Description

This function allows you to converting CpG site database file (text file type) to data frame for analysis. User can download CpG site database file from UCSC genome browser.

Usage

```
readCpGdb(cpgfile, select = TRUE)
```

Arguments

cpgfile	Full path of CpG site database file(text file format) from UCSC genome browser.
select	If you want to include small size (<300bps) CpG sites in output, that should be FALSE. TRUE is counterpart of that. Default is TRUE.

Examples

```
readCpGdb(cpgfile = ~/test/chicken_raw.cpg, select = FALSE)
```

readGFF — *Function for reading GFF file*

Description

This function allows you to converting a gff file to data frame for analysis. User can download GFF files from NCBI Refseq. Before using `extractFeatures` function, you should run this function.

Usage

```
readGFF(gffFile, nrows = -1)
```

Arguments

gffFile	Full path of target's annotation file (GFF format).
nrows	Do not modify negative value of this.

See Also

`extractFeatures`

Examples

```
readGFF(gffFile = ~/test/chicken.gff)
```

readRepeatdb — *Function for converting annotation file from UCSC to specific data frame*

Description

This function allows you to converting repeat database file (text file type) to data frame for analysis. User can download repeat database file from UCSC genome browser.

Usage

```
readRepeatdb(repeatfile, includeSimpleRepeats = FALSE,
  includeUnknownClass = FALSE)
```

Arguments

repeatfile	Full path of repeat database file(text file format) from UCSC genome browser.
includeSimpleRepeats	If you want to include simple repeats in output, that should be TRUE. FALSE is counterpart of that. Default is FALSE.
includeUnknownClass	If you want to include repeats that are included in unknown class, that should be TRUE. FALSE is counterpart of that. Default is FALSE.

Examples

```
readRepeatdb(repeatfile = ~/test/chicken_raw.rdb, includeSimpleRepeats = TRUE, includeUnknownClass = FALSE)
```

readTSSdb — *Function for converting annotation file from DBTSS to specific data frame*

Description

This function allows you to converting TSS database file (tab file type) to data frame for analysis. User can download TSS database file from DBTSS.

Usage

```
readTSSdb(inputfile)
```

Arguments

inputfile	Full path of transcription start site database file(text file format) from DBTSS.

Examples

```
readTSSdb(inputfile = ~/test/human_HEK293.tab)
```

Index