Package 'chromseq'

May 11, 2020

Type Package

Title Split Chromosome 'Fasta' File

Description
Chromosome files in the 'Fasta' format usually contain large sequences like human genome. Sometimes users have to split these chromosomes into different files according to their chromosome number. The 'chromseq' can help to handle this. So the selected chromosome sequence can be
used for downstream analysis like motif finding. Howard Y. Chang(2019) <doi:10.1038/s41587-019-0206-z>.

Version 0.1.3

License Artistic-2.0

Encoding UTF-8

LazyData true

RoxygenNote 7.0.2

NeedsCompilation no

URL https://github.com/MSQ-123/chromseq

BugReports https://github.com/MSQ-123/chromseq/issues

Depends R (>= 2.10)

Imports utils, base

Author Shaoqian Ma [aut, cre]

Maintainer Shaoqian Ma <897341109@qq.com>

Repository CRAN

Date/Publication 2020-05-11 15:20:17 UTC

R topics documented:

id 2
readToList 2
replaceText 3
sortList 4

splitChr 5
subFasID 6
tex 6
text 7

Index **8**

`id` *Sampled Fasta file of chromosome sequence from hg19 blacklist*

Description

This dataset is sampled from The hg19 blacklist. For splitting a chromosome Fasta file, sometimes the Fasta identifier is too complicated to manipulate. This data can be used to show how to simplify the Fasta identifier.

Usage

```
data(id)
```

Format

A character sequence with 20 elements

References

Satpathy A T, Granja J M, Yost K E, et al. (2019) Nature biotechnology 37,925–936. (PubMed)

Examples

```
data(id)
```

`readToList` *Make a list file from large chromosome Fasta file*

Description

Make a list file from large chromosome Fasta file

Usage

```
readToList(id = id, text = text, con = con)
```

Arguments

id	The id list made from subFasID function
text	Large character read in by readLines function from Fasta file
con	A connection object or a character string, the connection must refer to the same Fasta file as text

Value

Chromosome Fasta file in list format.

Examples

```
data("text")
id <- subFasID(text = text)
fil <- tempfile(fileext = ".data")
write(text,file = fil)
con0 <- file(fil, "r")
tex <- readToList(id,text = text,con = con0)
```

---

replaceText *Replace tedious chromosome identifier into simple format*

Description

Make the chromosome id starting with ">" into simple format like ">chr:1091194-1093520...",this is helpful for sorting the chromosome according to their number

Usage

```
replaceText(type = "text", input = input)
```

Arguments

type	This can be either "text" or "list", The previous is a large character containing each line of the Fasta file, the latter is a list in which each element contains a unit of Fasta file
input	The large character or list containing ids that need to be simplified

Value

The large character or list of Chromosome Fasta file with simplified id.

Author(s)

Shaoqian Ma

Examples

```
data("id")
simpleID<- replaceText(type = "text",input = id)
```

`sortList` *Sort the chromosome list according to the chromosome number*

Description

Sort the chromosome list according to the chromosome number

Usage

```
sortList(id = id, tex = tex, chrsig = "single")
```

Arguments

`id`	The identifier list of the Fasta file made by subFasID
`tex`	A chromosome Fasta file in list format made by readToList function
`chrsig`	The number of characters of the chromosome, either "single" or "double", the previous means a single character following "chr" in the Fasta identifier, the latter means two characters following "chr" in the Fasta identifier. eg."chr1,chrX,chrY,chrM" is "single";"chr10,chr11" is "double". If you want to obtain both "single" and "double" sorted list of chromosome, try "single" and "double" respectively

Value

The sorted chromosome Fasta file in list format.

Examples

```
data("tex")
data("text")
text<- replaceText(type = "text",input = text)
id <- subFasID(text = text)
tex2<- sortList(id=id,tex = tex,chrsig = "single")
tex3 <- sortList(id=id,tex = tex,chrsig = "double")
```

splitChr *Split all chromosomes from the sorted chromosome list*

Description

Split all chromosomes from the sorted chromosome list

Usage

```
splitChr(tex = tex, chr = chr, sex = FALSE, outdir = ".")
```

Arguments

tex	The sorted chromosome list made by sortList function.
chr	The chromosome number sequence, if the chromosome list is "single" which means a single character following "chr" in the Fasta identifier, be sure starting with 1 and ending with 9; if the chromosome list is "double" which means two characters following "chr" in the Fasta identifier, be sure that starting with 10 but the ending can be changed.
sex	Whether to output the sex chromosomes like X chromosome and Y chromosome.
outdir	The output directory.

Value

Write the splitted chromosome Fasta file to separated txt files according to the chromosome number.

Author(s)

Shaoqian Ma

Examples

```
data(tex)
data(text)
#Simplify the Fasta id
text<- replaceText(type = "text",input = text)
#Subtract id
id <- subFasID(text = text)
#Sort the fasta according to the chromosome number in id
tex2<- sortList(id=id,tex = tex,chrsig = "single")
tex3 <- sortList(id=id,tex = tex,chrsig = "double")
outdir <- tempdir()
#Output the results
splitChr(tex = tex2,chr=seq(1,9),sex = TRUE,outdir = outdir)
splitChr(tex = tex3,chr=seq(10,22),sex = FALSE,outdir = outdir)
```

subFasID *Subtract chromosome ids from Fasta file*

Description

Subtract chromosome ids from Fasta file

Usage

```
subFasID(text = text)
```

Arguments

text	Large character read by readLines from chromosome Fasta file.

Value

The id list of the Fasta file.

Examples

```
data("text")
text<- replaceText(type = "text",input = text)
id <- subFasID(text = text)
```

tex *Fasta file of chromosome sequence produced from sequence character*

Description

Data from "Three representative inter and intra-subspecific crosses reveal the genetic architecture of reproductive isolation in rice."

Usage

```
data(tex)
```

Format

A large list containing 62 elements.

References

Li, G. et al. (2017) The Plant Journal 92, 349–362. (PubMed)

Examples

```
data(tex)
```

`text` *Fasta file of chromosome sequence*

Description

A downsampled dataset containing the hg19 chromosome sequence from the hg19 blacklist. The hg19 blacklist is obtained from the supplementary dataset from "Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion." The dataset is sent to the UCSC Table Browser for obtaining the corresponding sequence file. The sequence file is processed with replaceText function to simplify the fasta id. To best illustate the usage, the sequence file is downsampled.

Usage

```
data(text)
```

Format

A character sequence with 2099 elements.

References

Satpathy A T, Granja J M, Yost K E, et al. (2019) Nature biotechnology 37, 925–936. (PubMed)

Examples

```
data(text)
```

Index

∗Topic **datasets**
id, 2
tex, 6
text, 7

id, 2

readToList, 2
replaceText, 3

sortList, 4
splitChr, 5
subFasID, 6

tex, 6
text, 7