Package 'STRAH'

March 4, 2019

Type Package

Title Short Tandem Repeats Analysis of Hotspot Zones

Version 0.1.0

Author Philipp Hermann [aut, cre],
Monika Heinzl [aut],
Angelika Heissl [ctb],
Irene Tiemann-Boege [ctb],
Andreas Futschik [ctb]

Maintainer Philipp Hermann <philipp.hermann@jku.at>

Description This package enables to search for short tandem repeats (STR) in a specified region of any genome. This analysis can be expanded such that several regions (chromosomes) are studied. These STRs can be grouped into hotspot as well as flanking regions of user specified width. Hotspots are defined by the double strand break maps from Pratto et al. (2014). Moreover, the user can also search for a specified motif in a DNAStringSet-object, or a fasta-file, or a specified region of any genome.

Imports Biostrings (>= 2.38.4),
BSgenome.Hsapiens.UCSC.hg19

Depends R (>= 2.10)

License MIT + file LICENSE

Encoding UTF-8

LazyData TRUE

BugReports https://github.com/PhHermann/STRAH

RoxygenNote 6.1.1

R topics documented:

chr6_1580213_1582559 *Sequence of human chromosome 6 from 1,580,213 until 1,582,559*

Description

A DNAStringSet object containing the data of human chromosome 6 starting with position 1,580,213 and ending at 1,582,559.

Usage

```
data (chr6_1580213_1582559)
```

Format

The data set is a DNAStringSet object containing one sequence of length 2346 nucleotides.

See Also

motif_detection

chr6_1581473_1586032 *Sequence of human chromosome 6 from 1,581,473 until 1,586,032*

Description

A DNAStringSet object containing the data of human chromosome 6 starting with position 1,581,473 and ending at 1,586,032.

Usage

```
data (chr6_1581473_1586032)
```

Format

The data set is a DNAStringSet object containing one sequence of length 4559 nucleotides.

See Also

motif_detection

dsb_map *Data of the DsbMap*

Description

A dataset containing the PRDM9-A type hotspots of Pratto et al. 2014.

Usage

```
data (dsb_map)
```

Format

The data set contains 37527 rows and 27 columns. We provide information on the most important columns (column nr, column name) hereafter:

1, chrom The chromosome under study

2, start Start coordinates of the hotspot

3, end End coordinates of the hotspot

4-8, (AA1,AA2,AB1,AB2,AC)_strength Strength of the corresponding PRDM9-type hotspot.

9-15, (AA1,AA2,AB1,AB2,AC)_hotspots Dummy coding whether these positions (start/end) define a hotspot of given PRDM9-type

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841 Pratto, F., et al. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211).

See Also

STR_analysis, STR_detection

dsb_map_chimp_full *Data of the DsbMap for chimpanzees*

Description

A dataset containing all translated PRDM9-A type hotspots of Pratto et al. 2014 for chimpanzees. Translation was done via liftOver of the UCSC Genome Browser from the human genome (hg19) to the chimpanzees genome (panTro5).

Usage

```
data (dsb_map_chimp_full)
```

Format

The data set contains 64078 rows and 3 columns. We provide information on the columns (column nr, column name) hereafter:

1, chrom The chromosome under study

2, start Start coordinates of the hotspot

3, end End coordinates of the hotspot

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841
Pratto, F., et al. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211).

See Also

STR_analysis, STR_detection

`getflank2` *Extract a specified region of the (human) genome*

Description

This function extracts a specified region of the human genome with corresponding start and end position of the region under study.

Usage

```
getflank2(species, chrs, start.position, end.position)
```

Arguments

`species`	The human genome (version 19) is default but an alternative genome can be provided. For chimpanzees the parameter has to be BSgenome.Ptroglodytes.UCSC.panTro5 (given that the data is installed).
`chrs`	A string reflecting the chromosome under study (starting with "chr" and adding either the integers from 1-22 or "X" respectively "Y"). This argument can also be a vector of strings to study several chromosomes.
`start.position`	An integer value reflecting the start position of the region to be analyzed. If set to `NA` the analysis starts from the beginning of the chromosome.
`end.position`	An integer value reflecting the end position of the region to be analyzed. If set to `NA` the analysis is performed until the end of the chromosome.

Value

The DNA-sequence of the region under study (defined by the chromosome, start position and end position) is returned.

Author(s)

Philipp Hermann, <philipp.hermann@jku.at>, Monika Heinzl, <monika.heinzl@edumail.at>
Angelika Heissl, Irene Tiemann-Boege, Andreas Futschik

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841

See Also

STR_analysis

Examples

```
getflank2(BSgenome.Hsapiens.UCSC.hg19::Hsapiens, "chr1", 1, 100000)
```

`motif_detection` *DNA-motif detection in a given DNAStringSet, a given DNA-sequence in fasta-format or a specified region of any genome*

Description

This function searches for a given "DNA-motif" in a DNA-sequence. The argument `seqName` can be either a DNAStringSet object or it refers to a fasta-file. Additionally, we provide the option to specify a species, a chromosome, a start, and a stop position for a region of any reference genome to be analyzed. By default a region of the human genome is analyzed. Optionally, one can also specify the number of mismatches of the DNA-motif and whether the reverse complement has to be searched.

Usage

```
motif_detection(seqName, chrs, start.position, end.position, motif,
  nr.mismatch = 0, reverse.comp = F, print.status = T,
  species = BSgenome.Hsapiens.UCSC.hg19::Hsapiens)
```

Arguments

`seqName`	A character string which can either be the name of a DNAStringSet object or a sequence name referring to a fasta-file to be analyzed. This argument can only be ignored if `chr` and `start.position` and `end.position` are specified.
`chrs`	A character string reflecting the chromosome under study (starting with "chr" and adding either the integers from 1-22 or "X" respectively "Y"). This argument can also be a vector of strings to study several chromosomes.
`start.position`	An integer value reflecting the start position of the region to be analyzed. If set to `NA` the analysis starts from the beginning of the chromosome.
`end.position`	An integer value reflecting the end position of the region to be analyzed. If set to `NA` the analysis is performed until the end of the chromosome.
`motif`	A character string reflecting the specified DNA-motif to be searched for in the DNA-sequence.

nr.mismatch	This integer specifies the number of allowed mismatches when searching for the specified DNA-motif.
reverse.comp	A logical value, by default FALSE, which enables to search the reverse complement of the sequence if set to TRUE.
print.status	A logical value reflecting whether the current status of the worked sequence (relative to the sequence length) is printed (TRUE) or not (FALSE).
species	The human genome (version 19) is default but an alternative genome can be provided. For chimpanzees the parameter has to be BSgenome.Ptroglodytes.UCSC.panTro5 (given that the data is installed).

Value

The output of the function is a list with the following content:

Species	The name of the species under study
Sequence Name	The name of the region under study
Reverse Complement	Indicator whether the reverse complement was searched
Number of Matches	The frequency of found DNA-motifs in the region under study
Start Positions of Matches	The start positions of the found DNA-motifs
Number of allowed Mismatches	The number of allowed mismatches when searching for the DNA-motif
Matched Segments	The list of the segments containing the DNA-motif

Author(s)

Philipp Hermann, <philipp.hermann@jku.at>, Monika Heinzl, <monika.heinzl@edumail.at>
Angelika Heissl, Irene Tiemann-Boege, Andreas Futschik

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841

See Also

getflank2

Examples

```
data(chr6_1580213_1582559)
motif_detection(seqName = chr6_1580213_1582559, chrs="",
start.position = NA, end.position = NA, motif = "CCNCCNTNNCCNC",
nr.mismatch = 1, reverse.comp = FALSE, print.status = TRUE)

motif_detection(seqName = "", chrs = "chr6",
start.position = 1580213, end.position = 1582559,
motif = "CCNCCNTNNCCNC", nr.mismatch = 1, reverse.comp = FALSE, print.status = FALSE)
```

```
# If you want to use the function with a different reference genome
# make your choice and install it before:
# source("http://bioconductor.org/biocLite.R")
# biocLite("BSgenome.Ptroglodytes.UCSC.panTro5")
# library(BSgenome.Ptroglodytes.UCSC.panTro5)
# motif_detection(seqName = "", chrs = "chr1", start.position =222339618, end.position = 222339660,
# motif = "A", nr.mismatch = 0, reverse.comp = FALSE,
# print.status = FALSE, species = BSgenome.Ptroglodytes.UCSC.panTro5)
```

STRAH *STRAH: A package to detect DNA-motifs or short tandem repeats (STRs) in reference genomes*

Description

The STRAH package provides functions to extract DNA sequence data of reference genomes, functions to search for DNA-motifs in a defined DNA-sequence or to detect short tandem repeats (STRs) of specified length, and to analyze detected STRs in the human genome by comparing them with double strand break hotspots.

STRAH functions:

getflank2 Extract a specified region of any reference genome (e.g. humans, chimpanzees, ...)

motif_detection DNA-motif detection of a given DNAStringSet, a given DNA-sequence in fasta-format or a specified region of any genome

STR_detection Detection of short tandem repeats (STRs) in either one or several regions of a given species

STR_analysis Analysis of detected short tandem repeats (STRs) in either one or several regions of a given species

STRAH data sets:

chr6_1580213_1582559 DNA-sequence of human chromosome 6 from 1,580,213 until 1,582,559

chr6_1581473_1586032 DNA-sequence of human chromosome 6 from 1,581,473 until 1,586,032

dsb_map Data of the double strand break map of Pratto et al. (2014).

dsb_map_chimp_full Translated data of the double strand break map of Pratto et al. (2014) for the chimpanzees genome.

Author(s)

Philipp Hermann, <philipp.hermann@jku.at> (Maintainer), Monika Heinzl, <monika.heinzl@edumail.at>
Angelika Heissl, Irene Tiemann-Boege, Andreas Futschik

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841
Pratto, F., et al. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211).

STR_analysis *Analysis of short tandem repeats (STRs) in a given region of any reference genome*

Description

This function separates detected short tandem repeats (STRs) into different zones. These zones are either the hotspot zone defined by the double strand break maps of Pratto et al. (2014) or adjacent flanking zones (greyzones) left and right of the hotspots of user specified lengths. The parameters of the regions under study can be directly given in the function arguments or read in via either a BED-file or a position matrix.

Usage

```
STR_analysis(nr.STRs = 10, nr.mismatch = 0, chrs = "", STR = "A",
  lens.grey = lens.grey, start.position = NA, end.position = NA,
  reverse.comp = FALSE, bed_file = "", pos_matrix = "",
  output_file = "", species = BSgenome.Hsapiens.UCSC.hg19::Hsapiens,
  dsb_map = STRAH::dsb_map)
```

Arguments

nr.STRs	An integer value reflecting the minimum length of STRs to be searched for.
nr.mismatch	An integer value reflecting the allowed number of mismatches of the short tandem repeats. By default it set to 0.
chrs	A string reflecting the chromosome under study (starting with "chr" and adding either the integers from 1-22 or "X" respectively "Y"). This argument can also be a vector of strings to study several chromosomes.
STR	A character string for the nucleotide to be searched for. By default one searches for poly-As, hence set to "A".
lens.grey	An integer value which is by default a vector of 6 integer values. These values represent the greyzones to be studied left and right from the hotspot regions.
start.position	An integer value reflecting the start position of the region to be analyzed. If set to NA the analysis starts from the beginning of the chromosome.
end.position	An integer value reflecting the end position of the region to be analyzed. If set to NA the analysis is performed until the end of the chromosome.
reverse.comp	A logical value by default FALSE. If set to TRUE then the reverse complement of the sequence is analyzed.
bed_file	A bed file containing the chromosomes, start and end positions of the region(s) that should be analyzed.
pos_matrix	A matrix or dataframe containing the chromosomes, start and end positions of the region(s) that should be analyzed.
output_file	The default is an empty string and does not save an output-file. The output will be saved if the parameter is changed to a user defined string excluding the extension (by default .bed).
species	The human genome (version 19) is default but an alternative genome can be provided. For chimpanzees the parameter has to be BSgenome.Ptroglodytes.UCSC.panTro5 (given that the data is installed).

`dsb_map` The DSB map of the human genome (version 19) is default but an alternative DSB map from a different genome can be provided. This parameter needs to be a data frame with at least 3 columns that contains the chromosome, start and end position of the DSB. The DSB map for chimpanzees is included in the package.

Value

The output of the function is a list with the following content:

`Sequence Name` The chromosome with the starting and end position of the region under study is provided.

`Reverse Complement`
An indicator whether the reverse complement was considered

`Number of allowed Mismatches`
The number of allowed mismatches is provided.

`Minimum Length` The minimum length of the STR to be extracted is provided.

`Number of Matches`
The total number of STR matches of the region is provided.

`Length of STR stretch in bp`
A vector containing the length of STRs per match is provided.

`Start positions`
The starting positions of the STRs are provided.

`Zone` The zones where the STR is found are provided. 1 reflects within a hotspot, the last integer reflects that it is outside, and the integers between these two reflect the given flanking regions starting with 2 as the next closest region to the hotspot.

A BED file with the chromosomes, start, and end position of the STRs, length of the STR stretch, the zone where the STR was found, and the specified region that was analyzed are given as columns.

Author(s)

Philipp Hermann, <philipp.hermann@jku.at>, Monika Heinzl, <monika.heinzl@edumail.at>
Angelika Heissl, Irene Tiemann-Boege, Andreas Futschik

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841
Pratto, F., et al. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211).

See Also

`getflank2`, `STR_detection`

Examples

```
STR_analysis(nr.STRs = 10, nr.mismatch = 0, chrs = "chr22", STR = "A", lens.grey = 0:5*1000,
start.position = 30000000, end.position = 40000000, reverse.comp = FALSE,
bed_file = "", pos_matrix = "", output_file = "",
species = BSgenome.Hsapiens.UCSC.hg19::Hsapiens, dsb_map = STRAH::dsb_map)
```

```
# If you want to use the function with a different reference genome
# make your choice and install it before:
# source("http://bioconductor.org/biocLite.R")
# biocLite("BSgenome.Ptroglodytes.UCSC.panTro5")
# library(BSgenome.Ptroglodytes.UCSC.panTro5)
# STR_analysis(nr.STRs = 10, nr.mismatch = 0, chrs = "chr22", STR = "A", lens.grey = 0:5*1000,
# start.position = 30000000, end.position = 40000000, reverse.comp = FALSE, bed_file = "",
# pos_matrix = "", output_file = "", species = BSgenome.Ptroglodytes.UCSC.panTro5,
# dsb_map = STRAH::dsb_map_chimp_full)
```

`STR_detection` *Detection of short tandem repeats (STRs) in a given region of any reference genome*

Description

This function searches for short tandem repeats (STRs) in a specified region of any reference genome. The parameters of the regions under study can be directly given in the function arguments or read in via either a BED-file or a position matrix. We recommend to search for STRs of minimum length 6. Options to save the output or usage of any reference genome are provided.

Usage

```
STR_detection(seqName = "", chrs = "", start.position = NA,
  end.position = NA, bed_file = "", pos_matrix = "", nr.STRs,
  nr.mismatch = 0, reverse.comp = F, STR = "A",
  species = BSgenome.Hsapiens.UCSC.hg19::Hsapiens,
  translated_regions = F, output_file = "")
```

Arguments

`seqName`	A character string which is the name of the given sequence file under study. Can also be set to "" in order to analyze an defined sequence from any reference genome such as the package BSgenome.Hsapiens.UCSC.hg19 for humans.
`chrs`	A string reflecting the chromosome under study (starting with "chr" and adding either the integers from 1-22 or "X" respectively "Y"). This argument can also be a vector of strings to study several chromosomes.
`start.position`	An integer value reflecting the start position of the region to be analyzed. If set to `NA` the analysis starts from the beginning of the chromosome.
`end.position`	An integer value reflecting the end position of the region to be analyzed. If set to `NA` the analysis is performed until the end of the chromosome.
`bed_file`	A bed file containing the chromosomes, start, and end positions of the region(s) that should be analyzed.
`pos_matrix`	A matrix or dataframe containing the chromosomes, start, and end positions of the region(s) that should be analyzed.
`nr.STRs`	An integer value as the minimum length of STRs to be detected.
`nr.mismatch`	An iInteger value reflecting the allowed number of mismatches of the short tandem repeats. By defaults set to 0.
`reverse.comp`	A logical value by default `FALSE`. If set to `TRUE` then the reverse complement of the sequence is analyzed.

`STR` A character string for the nucleotide to be searched for. By default one searches for poly-As, hence set to "A".

`species` The human genome (version 19) is default but an alternative genome can be provided. For chimpanzees the parameter has to be BSgenome.Ptroglodytes.UCSC.panTro5 (given that the data is installed).

`translated_regions`
A logical value by default `FALSE`. If set to `TRUE` then the function assumes that the parameters start.position and end.position were translated by some tool (e.g. Liftover) from one species to another. The untranslated and translated positions are included in the output.

`output_file` The default is an empty string and does not save an output-file. The output will be saved if the parameter is changed to a user defined string excluding the extension (by default .bed).

Value

The output of the function is a list with the following content:

`Sequence name` The chromosome with the starting and end position of the region under study is provided. If translated_region is set to `TRUE`, the interval will be the translated region.

`Sequence name (untranslated)`
Only if translated_region is set to `TRUE`, then the untranslated region (chromosome with the starting and end position) is provided.

`Reverse complement`
An indicator whether the reverse complement was considered

`Number of allowed mismatches`
The number of allowed mismatches is provided.

`Minimum length` The minimum length of the STR to be extracted is provided.

`Number of matches`
The total number of STR matches of the region is provided.

`Length of STR stretch in bp`
A vector containing the length of STRs per match is provided.

`Start positions`
The starting positions of the STRs are provided.

`Matched segments`
The matched segments of the STRs are provided.

A BED file with chromosomes, start and end position of the STRs, length of the STR stretch, the matched segments and the specified region (untranslated and translated) that was analyzed are given as columns.

Author(s)

Philipp Hermann, `<philipp.hermann@jku.at>`, Monika Heinzl, `<monika.heinzl@edumail.at>`
Angelika Heissl, Irene Tiemann-Boege, Andreas Futschik

References

Heissl, A., et al. (2018) Length asymmetry and heterozygosity strongly influences the evolution of poly-A microsatellites at meiotic recombination hotspots. doi: https://doi.org/10.1101/431841
Pratto, F., et al. (2014). Recombination initiation maps of individual human genomes. Science, 346(6211).

See Also

getflank2, STR_analysis

Examples

```
STR_detection(seqName = "", chrs = "chr22", start.position = 30000000, end.position = 40000000,
nr.STRs = 10, nr.mismatch = 0, reverse.comp = FALSE, STR = "A",
species=BSgenome.Hsapiens.UCSC.hg19::Hsapiens, translated_regions=FALSE, output_file = "")
# If you want to use the function with a different reference genome
# make your choice and install it before:
# source("http://bioconductor.org/biocLite.R")
# biocLite("BSgenome.Ptroglodytes.UCSC.panTro5")
# library(BSgenome.Ptroglodytes.UCSC.panTro5)
# STR_detection(seqName = "", chrs = "chr1", start.position =222339618, end.position = 222339660,
# nr.STRs = 10, nr.mismatch = 0, reverse.comp = FALSE, STR = "A",
# species = BSgenome.Ptroglodytes.UCSC.panTro5)
```

Index