Package 'CRISPRseek'

April 9, 2015

Type Package

Title Design of target-specific guide RNAs in CRISPR-Cas9, genome-editing systems

Version 1.4.2

Date 2014-10-09

Author Lihua Julie Zhu, Benjamin R. Holmes, Herve Pages, Neil Aronin and Michael Brodsky

Maintainer Lihua Julie Zhu <julie.zhu@umassmed.edu>

Depends R (>= 3.0.1), BiocGenerics, Biostrings, BSgenome, seqinr

biocViews GeneRegulation, SequenceMatching

Suggests RUnit, BiocStyle, BSgenome.Hsapiens.UCSC.hg19, TxDb.Hsapiens.UCSC.hg19.knownGene, org.Hs.eg.db

Description The package includes functions to find potential guide RNAs for input target sequences, optionally filter guide RNAs without restriction enzyme cut site, or without paired guide RNAs, genome-wide search for off-targets, score, rank, fetch flank sequence and indicate whether the target and off-targets are located in exon region or not. Potential guide RNAs are annotated with total score of the top5 and topN off-targets, detailed topN mismatch sites, restriction enzyme cut sites, and paired guide RNAs. This package leverages Biostrings and BSgenome packages.

License GPL (>= 2)

LazyLoad yes

R topics documented:

`CRISPRseek-package` *Design of target-specific guide RNAs (gRNAs) in CRISPR-Cas9, genome-editing systems*

Description

Design of target-specific gRNAs for the CRISPR-Cas9 system by automatically finding potential gRNAs (paired/not paired), with/without restriction enzyme cut site(s) in a given sequence, searching for off targets with user defined maximum number of mismatches, calculating score of each off target based on mismatch positions in the off target and a penalty weight matrix, filtering off targets with user-defined criteria, and annotating off targets with flank sequences, whether located in exon or not. Summary report is also generated with gRNAs ranked by total topN off target score, annotated with restriction enzyme cut sites, gRNA efficacy and possible paired gRNAs. Detailed paired gRNAs information and restriction enzyme cut sites are stored in separate files in the output directory specified by the user. In total, four tab delimited files are generated in the output directory: OfftargetAnalysis.xls (off target details), Summary.xls (gRNA summary), REcutDetails.xls (restriction enzyme cut sites of each gRNA), and pairedgRNAs.xls (potential paired gRNAs).

Details

Package:	CRISPRseek
Type:	Package
Version:	1.0
Date:	2013-10-04
License:	GPL (>= 2)

Function offTargetAnalysis integrates all steps of off target analysis into one function call

Author(s)

Lihua Julie Zhu and Michael Brodsky Maintainer: julie.zhu@umassmed.edu

References

Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM.CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol. 2013. 31(9):833-8 Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao & Feng Zhang. DNA targeting specificity of rNA-guided Cas9 nucleases. Nat Biotechnol. 2013. 31:827-834

See Also

offTargetAnalysis

Examples

```
    library(CRISPRseek)
    library("BSgenome.Hsapiens.UCSC.hg19")
    library(TxDb.Hsapiens.UCSC.hg19.knownGene)
    library(org.Hs.eg.db)
    outputDir <- getwd()
    inputFilePath <- system.file("extdata", "inputseq.fa", package = "CRISPRseek")
    REpatternFile <- system.file("extdata", "NEBenzymes.fa", package = "CRISPRseek")
######## Scenario 1. Target and off-target analysis for paired gRNAs with
######## one of the pairs overlap RE sites
    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly=TRUE,
        REpatternFile =REpatternFile,findPairedgRNAOnly=TRUE,
        BSgenomeName=Hsapiens, txdb=TxDb.Hsapiens.UCSC.hg19.knownGene,
        orgAnn = org.Hs.egSYMBOL,max.mismatch = 1, chromToSearch = "chrX",
        outputDir = outputDir,overwrite = TRUE)

######## Scenario 2. Target and off-target analysis for paired gRNAs with or
######## without RE sites
    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = FALSE,
        REpatternFile = REpatternFile,findPairedgRNAOnly = TRUE,
        BSgenomeName = Hsapiens, txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
        orgAnn = org.Hs.egSYMBOL,max.mismatch = 1, chromToSearch = "chrX",
        outputDir = outputDir, overwrite = TRUE)

######## Scenario 3. Target and off-target analysis for gRNAs overlap RE sites

    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = TRUE,
        REpatternFile = REpatternFile,findPairedgRNAOnly = FALSE,
        BSgenomeName = Hsapiens, txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
        orgAnn = org.Hs.egSYMBOL, max.mismatch = 1, chromToSearch = "chrX",
        outputDir = outputDir, overwrite = TRUE)

######## Scenario 4. Off-target analysis for all potential gRNAs, this will
########be the slowest among the aforementioned scenarios.

    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = FALSE,
        REpatternFile = REpatternFile,findPairedgRNAOnly = FALSE,
        BSgenomeName = Hsapiens, txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
         orgAnn = org.Hs.egSYMBOL, max.mismatch = 1, chromToSearch = "chrX",
        outputDir = outputDir,overwrite = TRUE)

######## Scenario 5. Target and off-target analysis for gRNAs input by user.
    gRNAFilePath <- system.file("extdata", "testHsap_GATA1_ex2_gRNA1.fa",
        package="CRISPRseek")
    results <- offTargetAnalysis(inputFilePath = gRNAFilePath, findgRNAs = FALSE,
        findgRNAsWithREcutOnly = FALSE, REpatternFile = REpatternFile,
        findPairedgRNAOnly = FALSE, BSgenomeName = Hsapiens,
        txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
```

```
         orgAnn = org.Hs.egSYMBOL, max.mismatch = 1, chromToSearch = "chrX",
        outputDir = outputDir, overwrite = TRUE)

####### Scenario 6. Quick gRNA finding  without target and off-target analysis
    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = TRUE,
        REpatternFile = REpatternFile,findPairedgRNAOnly = TRUE,
        chromToSearch = "", outputDir = outputDir, overwrite = TRUE)

####### Scenario 7. Quick gRNA finding  with gRNA efficacy analysis
    results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = TRUE,
        REpatternFile = REpatternFile,findPairedgRNAOnly = TRUE,
BSgenomeName = Hsapiens, annotateExon = FALSE,
        max.mismatch = 0, outputDir = outputDir, overwrite = TRUE)
```

`buildFeatureVectorForScoring` *Build feature vectors*

Description

Build feature vectors for calculating scores of off targets

Usage

```
buildFeatureVectorForScoring(hits, gRNA.size = 20, canonical.PAM = "NGG")
```

Arguments

`hits`	a data frame generated from searchHits, which contains IsMismatch.posX (Indicator variable indicating whether this position X is mismatch or not, 1 means yes and 0 means not, X = 1- gRNA.size) representing all positions in the guide RNA, abbreviated as gRNA),strand (strand of the off target, + for plus and - for minus strand), chrom (chromosome of the off target), chromStart (start position of the off target), chromEnd (end position of the off target),name (gRNA name), gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (set to 100, and will be calculated in getOfftargetScore)
`gRNA.size`	gRNA size, default 20
`canonical.PAM`	Canonical PAM, default NGG

Value

A data frame with hits plus features used for calculating scores and for generating report, including IsMismatch.posX (Indicator variable indicating whether this position X is mismatch or not, 1 means yes and 0 means not, X = 1- gRNA.size) representing all positions in the gRNA), strand

(strand of the off target, + for plus and - for minus strand),chrom (chromosome of the off target), chromStart (start position of the off target),chromEnd (end position of the off target),name (gRNA name), gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (score of the off target), mismatche.distance2PAM (a comma separated distances of all mismatches to PAM, e.g., 14,11 means one mismatch is 14 bp away from PAM and the other mismatch is 11 bp away from PAM), alignment (alignment between gRNA and off target, e.g., ...G..C..... means that this off target aligns with gRNA except that G and C are mismatches), NGG (this off target contains canonical PAM or not, 1 for yes and 0 for no) mean.neighbor.distance.mismatch (mean distance between neighboring mismatches)

Author(s)

Lihua Julie Zhu

See Also

offTargetAnalysis

Examples

```
hitsFile <-  system.file("extdata", "hits.txt", package = "CRISPRseek")
hits <- read.table(hitsFile, sep= "\t", header = TRUE,
    stringsAsFactors = FALSE)
buildFeatureVectorForScoring(hits)
```

`calculategRNAEfficiency` *Calculate gRNA Efficiency*

Description

Calculate gRNA Efficiency for a given set of sequences and feature weight matrix

Usage

```
calculategRNAEfficiency <- function(extendedSequence, baseBeforegRNA, featureWeightMatrix, gRNA.size
```

Arguments

`extendedSequence`	Sequences containing gRNA plus PAM plus flanking sequences. Each sequence should be long enough for building features specified in the featureWeightMatrix
`baseBeforegRNA`	Number of bases before gRNA used for calculating gRNA efficiency, default 4

`featureWeightMatrix`
: a data frame with the first column containing significant features and the second column containing the weight of corresponding features. In the following example, DoenchNBT2014 weight matrix is used. Briefly, features include INTERCEPT,GC_LOW (penalty for low GC content in the gRNA sequence), GC_HIGH (penalty for high GC content in the gRNA sequence), G02 (means G at second position of the extendedSequence), GT02 (means GT di-nucleotides starting at 2nd position of the extendedSequence). To understand how is the feature weight matrix is identified, or how to use alternative feature weight matrix file, please see Doench et al., 2014 for details.

`gRNA.size`
: The size of the gRNA, default 20

Value

DNAStringSet consists of potential gRNAs that can be input to filtergRNAs function directly

Author(s)

Lihua Julie Zhu

References

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014 Sep 3. doi: 10.1038 nbt.3026 http://www.broadinstitute.org/rnai/public/analysis-tools/sgrna-design

See Also

offTargetAnalysis

Examples

```
  extendedSequence <- c("TGGATTGTATAATCAGCATGGATTTGGAAC",
"TCAACGAGGATATTCTCAGGCTTCAGGTCC",
"GTTACCTGAATTTGACCTGCTCGGAGGTAA",
"CTTGGTGTGGCTTCCTTTAAGACATGGAGC",
"CATACAGGCATTGAAGAAGAATTTAGGCCT",
"AGTACTATACATTTGGCTTAGATTTGGCGG",
"TTTTCCAGATAGCCGATCTTGGTGTGGCTT",
"AAGAAGGGAACTATTCGCTGGTGATGGAGT"
)
featureWeightMatrixFile <- system.file("extdata", "DoenchNBT2014.csv",
package = "CRISPRseek")
featureWeightMatrix <- read.csv(featureWeightMatrixFile, header=TRUE)
calculategRNAEfficiency(extendedSequence, baseBeforegRNA = 4,
featureWeightMatrix, gRNA.size = 20)
```

compare2Sequences *Compare 2 input sequences/sequence sets for possible guide RNAs (gRNAs)*

Description

Generate all possible guide RNAs (gRNAs) for two input sequences, or two sets of sequences and generate scores for potential off-targets in the other sequence.

Usage

```
compare2Sequences(inputFile1Path, inputFile2Path,
    inputNames=c("Seq1", "Seq2"),
    format = "fasta", findgRNAsWithREcutOnly = FALSE,
    searchDirection=c("both","1to2", "2to1"),
   REpatternFile=system.file("extdata", "NEBenzymes.fa", package = "CRISPRseek"),
    minREpatternSize = 6,
    overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = FALSE,
    min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3,
    gRNA.size = 20, PAM = "NGG", PAM.pattern = "N[A|G]G$", max.mismatch = 3,
    outputDir,
    weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445,
    0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583),
    overwrite = FALSE)
```

Arguments

inputFile1Path	Sequence input file 1 path that contains one of the two sequences to be searched for potential gRNAs
inputFile2Path	Sequence input file 2 path that contains one of the two sequences to be searched for potential gRNAs
inputNames	Name of the input sequences when inputFile1Path and inputFile2Path are DNAStringSet instead of file path
format	Format of the input file, fasta and fastq are supported, default fasta
findgRNAsWithREcutOnly	Indicate whether to find gRNAs overlap with restriction enzyme recognition pattern
searchDirection	Indicate whether perfrom gRNA in both sequences and off-target search against each other (both) or search gRNA in input1 and off-target analysis in input2 (1to2), or vice versa (2to1)
REpatternFile	File path containing restriction enzyme cut patters
minREpatternSize	Minimum restriction enzyme recognition pattern length required for the enzyme pattern to be searched for, default 6

`overlap.gRNA.positions`
: The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18

`findPairedgRNAOnly`
: Choose whether to only search for paired gRNAs in such an orientation that the first one is on minus strand called reverse gRNA and the second one is on plus strand called forward gRNA. TRUE or FALSE, default FALSE

`min.gap`
: Minimum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 0

`max.gap`
: Maximum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 20

`gRNA.name.prefix`
: The prefix used when assign name to found gRNAs, default gRNA, short for guided RNA.

`PAM.size`
: PAM length, default 3

`gRNA.size`
: The size of the gRNA, default 20

`PAM`
: PAM sequence after the gRNA, default NGG

`PAM.pattern`
: Regular expression of PAM, default N[A|G]G$

`max.mismatch`
: Maximum mismatch allowed to search the off targets in the other sequence, default 3

`outputDir`
: the directory where the sequence comparison results will be written to

`weights`
: numeric vector size of gRNA length, default c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583) which is used in Hsu et al., 2013 cited in the reference section

`overwrite`
: overwrite the existing files in the output directory or not, default TRUE

Value

Return a data frame with all potential gRNAs from both sequences. In addition, a tab delimited file scoresFor2InputSequences.xls is also saved in the outputDir, sorted by scoreDiff descending.

`name`
: name of the gRNA

`gRNAPlusPAM`
: gRNA plus PAM sequence

`targetInSeq1`
: target/off-target sequence including PAM in the 1st input sequence file

`targetInSeq2`
: target/off-target sequence incuding PAM in the 2nd input sequence file

`guideAlignment2Offtarget`
: alignment of gRNA to the other input sequence (off-target sequence)

`offTargetStrand`
: strand of the other sequence (off-target sequence) the gRNA align to

`scoreForSeq1`
: score for the target sequence in the 1st input sequence file

`scoreForSeq2`
: score for the target sequence in the 1st input sequence file

`mismatch.distance2PAM`
: distances of mismatch to PAM, e.g., 14 means the mismatch is 14 bp away from PAM

n.mismatch	number of mismatches between the off-target and the gRNA
targetSeqName	the name of the input sequence where the target sequence is located
scoreDiff	scoreForSeq1 - scoreForSeq2

Author(s)

Lihua Julie Zhu

References

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao & Feng Zhang (2013) DNA targeting specificity of rNA-guided Cas9 nucleases. Nature Biotechnology 31:827-834

See Also

CRISPRseek

Examples

```
library(CRISPRseek)
inputFile1Path <- system.file("extdata", "rs362331T.fa",
        package = "CRISPRseek")
inputFile2Path <- system.file("extdata", "rs362331C.fa",
        package = "CRISPRseek")
REpatternFile <- system.file("extdata", "NEBenzymes.fa",
        package = "CRISPRseek")
seqs <- compare2Sequences(inputFile1Path, inputFile2Path,
    outputDir = getwd(),
    REpatternFile = REpatternFile, overwrite = TRUE)
```

---

filtergRNAs *Filter gRNAs*

Description

Filter gRNAs containing restriction enzyme cut site

Usage

```
filtergRNAs(all.gRNAs, pairOutputFile = "",
    findgRNAsWithREcutOnly = FALSE,
    REpatternFile = system.file("extdata", "NEBenzymes.fa",
        package = "CRISPRseek"), format = "fasta",
    minREpatternSize = 6, overlap.gRNA.positions = c(17, 18))
```

Arguments

all.gRNAs	gRNAs as DNAStringSet, such as the output from findgRNAs
pairOutputFile	File path with paired gRNAs
findgRNAsWithREcutOnly	Indicate whether to find gRNAs overlap with restriction enzyme recognition pattern
REpatternFile	File path containing restriction enzyme cut patters
format	Format of the REpatternFile, default as fasta
minREpatternSize	Minimum restriction enzyme recognition pattern length required for the enzyme pattern to be searched for, default 6
overlap.gRNA.positions	The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18

Value

gRNAs.withRE	gRNAs as DNAStringSet that passed the filter criteria
gRNAREcutDetails	a data frame that contains a set of gRNAs annotated with restriction enzyme cut details

Author(s)

Lihua Julie Zhu

See Also

offTargetAnalysis

Examples

```
all.gRNAs <- findgRNAs(
    inputFilePath = system.file("extdata", "inputseq.fa",
    package = "CRISPRseek"),
    pairOutputFile = "testpairedgRNAs.xls",
    findPairedgRNAOnly = TRUE)

gRNAs.RE <- filtergRNAs(all.gRNAs = all.gRNAs,
    pairOutputFile = "testpairedgRNAs.xls",
    REpatternFile = system.file("extdata", "NEBenzymes.fa",
    package = "CRISPRseek"))

gRNAs  <- gRNAs.RE$gRNAs.withRE
restriction.enzyme.cut.sites <- gRNAs.RE$gRNAREcutDetails
```

filterOffTarget *filter off targets and generate reports.*

Description

filter off targets that meet the criteria set by users such as minimum score, topN. In addition, off target was annotated with flank sequence, gRNA cleavage efficiency and whether it is inside an exon or not if fetchSequence is set to TRUE and annotateExon is set to TRUE

Usage

```
filterOffTarget(scores, min.score = 0.5, topN = 100,
    topN.OfftargetTotalScore = 10,
    annotateExon = TRUE, txdb, orgAnn, outputDir, oneFilePergRNA = FALSE,
    fetchSequence = TRUE, upstream = 200, downstream = 200, BSgenomeName,
    baseBeforegRNA = 4, baseAfterPAM = 3,
    featureWeightMatrixFile = system.file("extdata", "DoenchNBT2014.csv",
package = "CRISPRseek"))
```

Arguments

scores	a data frame output from getOfftargetScore. It contains strand (strand of the off target, + for plus and - for minus strand), chrom (chromosome of the off target), chromStart (start position of the off target), chromEnd (end position of the off target),name (gRNA name),gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (score of the off target), mismatche.distance2PAM (a comma separated distances of all mismatches to PAM, e.g., 14,11 means one mismatch is 14 bp away from PAM and the other mismatch is 11 bp away from PAM), alignment (alignment between gRNA and off target, e.g., ...G..C..... means that this off target aligns with gRNA except that G and C are mismatches),NGG (this off target contains canonical PAM or not, 1 for yes and 0 for no) mean.neighbor.distance.mismatch (mean distance between neighboring mismatches)
min.score	minimum score of an off target to included in the final output, default 0.5
topN	top N off targets to be included in the final output, default 100
topN.OfftargetTotalScore	top N off target used to calculate the total off target score, default 10
annotateExon	Choose whether or not to indicate whether the off target is inside an exon or not, default TRUE
txdb	TxDb object, for creating and using TxDb object, please refer to GenomicFeatures package. For a list of existing TxDb object, please search for annotation package starting with Txdb at http://www.bioconductor.org/packages/release/BiocViews.html#___Annota such as TxDb.Rnorvegicus.UCSC.rn5.refGene for rat, TxDb.Mmusculus.UCSC.mm10.knownGene

	for mouse, TxDb.Hsapiens.UCSC.hg19.knownGene for human, TxDb.Dmelanogaster.UCSC.dm3.ensGe for Drosophila and TxDb.Celegans.UCSC.ce6.ensGene for C.elegans
orgAnn	organism annotation dataset such as org.Hs.egSYMBOL for human
outputDir	the directory where the off target analysis and reports will be written to
oneFilePergRNA	write to one file for each gRNA or not, default to FALSE
fetchSequence	Fetch flank sequence of off target or not, default TRUE
upstream	upstream offset from the off target start, default 200
downstream	downstream offset from the off target end, default 200
BSgenomeName	BSgenome object. Please refer to available.genomes in BSgenome package. For example, BSgenome.Hsapiens.UCSC.hg19 for hg19, BSgenome.Mmusculus.UCSC.mm10 for mm10, BSgenome.Celegans.UCSC.ce6 for ce6, BSgenome.Rnorvegicus.UCSC.rn5 for rn5, and BSgenome.Dmelanogaster.UCSC.dm3 for dm3
baseBeforegRNA	Number of bases before gRNA used for calculating gRNA efficiency, default 4
baseAfterPAM	Number of bases after PAM used for calculating gRNA efficiency, default 3
featureWeightMatrixFile	Feature weight matrix file used for calculating gRNA efficiency. By default DoenchNBT2014 weight matrix is used. To use alternative weight matrix file, please input a csv file with first column containing significant features and the second column containing the corresponding weights for the features. Please see Doench et al., 2014 for details.

Value

offtargets	a data frame with off target analysis results
summary	a data frame with summary of the off target analysis results

Author(s)

Lihua Julie Zhu

References

Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014 Sep 3. doi: 10.1038 nbt.3026

See Also

offTargetAnalysis

Examples

```
library(CRISPRseek)
library("BSgenome.Hsapiens.UCSC.hg19")
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)
hitsFile <-  system.file("extdata", "hits.txt", package="CRISPRseek")
```

```
hits <- read.table(hitsFile, sep = "\t", header = TRUE,
    stringsAsFactors = FALSE)
featureVectors <- buildFeatureVectorForScoring(hits)
scores <- getOfftargetScore(featureVectors)
outputDir <- getwd()
results <- filterOffTarget(scores, BSgenomeName = Hsapiens,
    txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
     orgAnn = org.Hs.egSYMBOL, outputDir = outputDir,
    min.score = 0.1, topN = 10, topN.OfftargetTotalScore = 5)
results$offtargets
results$summary
```

`findgRNAs` *Find potential gRNAs*

Description

Find potential gRNAs for an input file containing sequences in fasta format

Usage

```
findgRNAs(inputFilePath, format = "fasta", PAM = "NGG", PAM.size = 3,
   findPairedgRNAOnly = FALSE, gRNA.pattern = "", gRNA.size = 20, min.gap = 0, max.gap = 20,
    pairOutputFile, name.prefix = "gRNA")
```

Arguments

`inputFilePath`	Sequence input file path or a DNAStringSet object that contains sequences to be searched for potential gRNAs
`format`	Format of the input file, fasta and fastq are supported, default fasta
`PAM`	protospacer-adjacent motif (PAM) sequence after the gRNA, default NGG
`PAM.size`	PAM length, default 3
`findPairedgRNAOnly`	Choose whether to only search for paired gRNAs in such an orientation that the first one is on minus strand called reverse gRNA and the second one is on plus strand called forward gRNA. TRUE or FALSE, default FALSE
`gRNA.pattern`	Regular expression or IUPAC Extended Genetic Alphabet to represent gRNA pattern, default is no restriction. To specify that the gRNA must start with GG for example, then set it to ^GG. Please see help(translatePattern) for a list of IUPAC Extended Genetic Alphabet.
`gRNA.size`	The size of the gRNA, default 20
`min.gap`	Minimum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 0
`max.gap`	Maximum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 20
`pairOutputFile`	The output file for writing paired gRNA information to
`name.prefix`	The prefix used when assign name to found gRNAs, default gRNA, short for guided RNA.

Details

If users already has a fasta file that contains a set of potential gRNAs, then users can call filergRNAs directly although the easiest way is to call the one-stop-shopping function OffTargetAnalysis with findgRNAs set to FALSE.

Value

DNAStringSet consists of potential gRNAs that can be input to filtergRNAs function directly

Note

If the input sequence file contains multiple >300 bp sequences, suggest create one input file for each sequence and run the OffTargetAnalysis separately.

Author(s)

Lihua Julie Zhu

See Also

offTargetAnalysis

Examples

```
findgRNAs(inputFilePath = system.file("extdata",
    "inputseq.fa", package = "CRISPRseek"),
    pairOutputFile = "testpairedgRNAs.xls",
    findPairedgRNAOnly = TRUE)
```

---

getOfftargetScore *Calculate score for each off target*

Description

Calculate score for each off target with given feature vectors and weights vector

Usage

```
getOfftargetScore(featureVectors,
    weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508,
    0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583))
```

Arguments

`featureVectors`	a data frame generated from buildFeatureVectorForScoring. It contains IsMismatch.posX (Indicator variable indicating whether this position X is mismatch or not, 1 means yes and 0 means not, X = 1- gRNA.size) representing all positions in the gRNA), strand (strand of the off target, + for plus and - for minus strand), chrom (chromosome of the off target), chromStart (start position of the off target), chromEnd (end position of the off target), name (gRNA name),gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (score of the off target), mismatche.distance2PAM (a comma separated distances of all mismatches to PAM, e.g., 14,11 means one mismatch is 14 bp away from PAM and the other mismatch is 11 bp away from PAM), alignment (alignment between gRNA and off target, e.g., ...G..C..... means that this off target aligns with gRNA except that G and C are mismatches),NGG (this off target contains canonical PAM or not, 1 for yes and 0 for no) mean.neighbor.distance.mismatch (mean distance between neighboring mismatches)
`weights`	a numeric vector size of gRNA length, default c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583) which is used in Hsu et al., 2013 cited in the reference section

Details

score is calculated using the weights and algorithm by Hsu et al., 2013 cited in the reference section

Value

a data frame containing strand (strand of the match, + for plus and - for minus strand), chrom (chromosome of the off target), chromStart (start position of the off target), chromEnd (end position of the off target),name (gRNA name), gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (score of the off target), mismatche.distance2PAM (a comma separated distances of all mismatches to PAM, e.g., 14,11 means one mismatch is 14 bp away from PAM and the other mismatch is 11 bp away from PAM), alignment (alignment between gRNA and off target, e.g., ...G..C..... means that this off target aligns with gRNA except that G and C are mismatches), NGG (this off target contains canonical PAM or not, 1 for yes and 0 for no) mean.neighbor.distance.mismatch (mean distance between neighboring mismatches)

Author(s)

Lihua Julie Zhu

References

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini,

Gang Bao & Feng Zhang (2013) DNA targeting specificity of rNA-guided Cas9 nucleases. Nature Biotechnology 31:827-834

See Also

offTargetAnalysis

Examples

```
hitsFile <-  system.file("extdata", "hits.txt",
    package = "CRISPRseek")
hits <- read.table(hitsFile, sep = "\t", header = TRUE,
    stringsAsFactors = FALSE)
featureVectors <- buildFeatureVectorForScoring(hits)
getOfftargetScore(featureVectors)
```

`offTargetAnalysis` *Design of target-specific guide RNAs for CRISPR-Cas9 system in one function*

Description

Design of target-specific guide RNAs (gRNAs) for CRISPR-Cas9 system by automatically calling findgRNAs, filtergRNAs, searchHits, buildFeatureVectorForScoring, getOfftargetScore, filterOff-target, calculating gRNA cleavage efficiency and generate reports.

Usage

```
offTargetAnalysis(inputFilePath, format = "fasta", gRNAoutputName, findgRNAs = TRUE,
    exportAllgRNAs = c("all", "fasta", "genbank", "no"),
    findgRNAsWithREcutOnly = TRUE,
    REpatternFile = system.file("extdata", "NEBenzymes.fa",
        package = "CRISPRseek"), minREpatternSize = 6,
    overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE,
    min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3,
    gRNA.size = 20,  PAM = "NGG",  BSgenomeName, chromToSearch = "all",
    max.mismatch = 3, PAM.pattern = "N[A|G]G$", gRNA.pattern = "",
    min.score = 0.5, topN = 100,
    topN.OfftargetTotalScore = 10, annotateExon = TRUE,
   txdb, orgAnn, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200,
    weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508,
    0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583),
    baseBeforegRNA = 4, baseAfterPAM = 3,
    featureWeightMatrixFile = system.file("extdata", "DoenchNBT2014.csv",
package = "CRISPRseek"), useScore = TRUE,
    overwrite = FALSE)
```

Arguments

inputFilePath	Sequence input file path or a DNAStringSet object that contains sequences to be searched for potential gRNAs
format	Format of the input file, fasta and fastq are supported, default fasta
gRNAoutputName	Specify the name of the gRNA outupt file when inputFilePath is DNAStringSet object instead of file path
findgRNAs	Indicate whether to find gRNAs from the sequences in the input file or skip the step of finding gRNAs, default TRUE. Set it to FALSE if the input file contains user selected gRNAs plus PAM already.
exportAllgRNAs	Indicate whether to output all potential gRNAs to a file in fasta format, genbank format or both. Default to both.
findgRNAsWithREcutOnly	Indicate whether to find gRNAs overlap with restriction enzyme recognition pattern
REpatternFile	File path containing restriction enzyme cut patterns
minREpatternSize	Minimum restriction enzyme recognition pattern length required for the enzyme pattern to be searched for, default 6
overlap.gRNA.positions	The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18
findPairedgRNAOnly	Choose whether to only search for paired gRNAs in such an orientation that the first one is on minus strand called reverse gRNA and the second one is on plus strand called forward gRNA. TRUE or FALSE, default FALSE
min.gap	Minimum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 0
max.gap	Maximum distance between two oppositely oriented gRNAs to be valid paired gRNAs. Default 20
gRNA.name.prefix	The prefix used when assign name to found gRNAs, default gRNA, short for guided RNA.
PAM.size	PAM length, default 3
gRNA.size	The size of the gRNA, default 20
PAM	PAM sequence after the gRNA, default NGG
BSgenomeName	BSgenome object. Please refer to available.genomes in BSgenome package. For example, BSgenome.Hsapiens.UCSC.hg19 for hg19, BSgenome.Mmusculus.UCSC.mm10 for mm10, BSgenome.Celegans.UCSC.ce6 for ce6, BSgenome.Rnorvegicus.UCSC.rn5 for rn5, BSgenome.Drerio.UCSC.danRer7 for Zv9, and BSgenome.Dmelanogaster.UCSC.dm3 for dm3
chromToSearch	Specify the chromosome to search, default to all, meaning search all chromosomes. For example, chrX indicates searching for matching in chromosome X only

`max.mismatch` Maximum mismatch allowed in off target search, default 3. Warning: will be considerably slower if set > 3

`PAM.pattern` Regular expression of protospacer-adjacent motif (PAM), default N[A|G]G$

`gRNA.pattern` Regular expression or IUPAC Extended Genetic Alphabet to represent gRNA pattern, default is no restriction. To specify that the gRNA must start with GG for example, then set it to ^GG. Please see help(translatePattern) for a list of IUPAC Extended Genetic Alphabet.

`min.score` minimum score of an off target to included in the final output, default 0.5

`topN` top N off targets to be included in the final output, default 100

`topN.OfftargetTotalScore` top N off target used to calculate the total off target score, default 10

`annotateExon` Choose whether or not to indicate whether the off target is inside an exon or not, default TRUE

`txdb` TxDb object, for creating and using TxDb object, please refer to GenomicFeatures package. For a list of existing TxDb object, please search for annotation package starting with Txdb at http://www.bioconductor.org/packages/release/BiocViews.html#___Annota such as TxDb.Rnorvegicus.UCSC.rn5.refGene for rat, TxDb.Mmusculus.UCSC.mm10.knownGene for mouse, TxDb.Hsapiens.UCSC.hg19.knownGene for human, TxDb.Dmelanogaster.UCSC.dm3.ensGe for Drosophila and TxDb.Celegans.UCSC.ce6.ensGene for C.elegans

`orgAnn` organism annotation dataset such as org.Hs.egSYMBOL from org.Hs.eg.db package for human

`outputDir` the directory where the off target analysis and reports will be written to

`fetchSequence` Fetch flank sequence of off target or not, default TRUE

`upstream` upstream offset from the off target start, default 200

`downstream` downstream offset from the off target end, default 200

`weights` a numeric vector size of gRNA length, default c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583) which is used in Hsu et al., 2013 cited in the reference section

`baseBeforegRNA` Number of bases before gRNA used for calculating gRNA efficiency, default 4

`baseAfterPAM` Number of bases after PAM used for calculating gRNA efficiency, default 3

`featureWeightMatrixFile` Feature weight matrix file used for calculating gRNA efficiency. By default DoenchNBT2014 weight matrix is used. To use alternative weight matrix file, please input a csv file with first column containing significant features and the second column containing the corresponding weights for the features. Please see Doench et al., 2014 for details.

`useScore` Default TRUE, display in gray scale with the darkness indicating the gRNA efficacy. The taller bar shows the Cas9 cutting site. If set to False, efficacy will not show. Instead, gRNAs in plus strand will be colored red and gRNAs in negative strand will be colored green.

`overwrite` overwrite the existing files in the output directory or not, default FALSE

Value

Four tab delimited files are generated in the output directory: OfftargetAnalysis.xls (detailed information of off targets), Summary.xls (summary of the gRNAs), REcutDetails.xls (restriction enzyme cut sites of each gRNA), and pairedgRNAs.xls (potential paired gRNAs)

Author(s)

Lihua Julie Zhu

References

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, Thomas J Cradick, Luciano A Marraffini, Gang Bao & Feng Zhang (2013) DNA targeting specificity of rNA-guided Cas9 nucleases. Nature Biotechnology 31:827-834 Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol. 2014 Sep 3. doi: 10.1038 nbt.3026

See Also

CRISPRseek

Examples

```
library(CRISPRseek)
library("BSgenome.Hsapiens.UCSC.hg19")
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
library(org.Hs.eg.db)
outputDir <- getwd()
inputFilePath <- system.file("extdata", "inputseq.fa",
            package = "CRISPRseek")
REpatternFile <- system.file("extdata", "NEBenzymes.fa",
            package = "CRISPRseek")
results <- offTargetAnalysis(inputFilePath, findgRNAsWithREcutOnly = TRUE,
            REpatternFile = REpatternFile, findPairedgRNAOnly = FALSE,
            BSgenomeName = Hsapiens, chromToSearch = "chrX",
            txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
    orgAnn = org.Hs.egSYMBOL, max.mismatch = 1,
            outputDir = outputDir, overwrite = TRUE)
```

---

searchHits *Search for off targets*

Description

Search for off targets for given gRNAs, BSgenome and maximum mismatches

Usage

```
searchHits(gRNAs, BSgenomeName, chromToSearch = "all", max.mismatch = 3,
    PAM.size = 3, gRNA.size = 20, PAM = "N[A|G]G$")
```

Arguments

gRNAs	DNAStringSet object containing a set of gRNAs. Please note the sequences must contain PAM appended after gRNAs, e.g., ATCGAAATTCGAGCCAATCCCGG where ATCGAAATTCGAGCCAATCC is the gRNA and CGG is the PAM
BSgenomeName	BSgenome object. Please refer to available.genomes in BSgenome package. For example, BSgenome.Hsapiens.UCSC.hg19 for hg19, BSgenome.Mmusculus.UCSC.mm10 for mm10, BSgenome.Celegans.UCSC.ce6 for ce6, BSgenome.Rnorvegicus.UCSC.rn5 for rn5, and BSgenome.Dmelanogaster.UCSC.dm3 for dm3
chromToSearch	Specify the chromosome to search, default to all, meaning search all chromosomes. For example, chrX indicates searching for matching in chromosome X only
max.mismatch	Maximum mismatch allowed in off target search, default 3. Warning: will be considerably slower if it is set to greater than 3
PAM.size	Size of PAM, default 3
gRNA.size	Size of gRNA, default 20
PAM	Regular expression of PAM, default N[A|G]G$

Value

a data frame contains IsMismatch.posX (indicator variable indicating whether this position X is mismatch or not, 1 means yes and 0 means not, X = 1 to gRNA.size) representing all positions in the gRNA),strand (strand of the match, + for plus and - for minus strand), chrom (chromosome of the off target), chromStart (start position of the off target), chromEnd (end position of the off target),name (gRNA name), gRNAPlusPAM (gRNA sequence with PAM sequence concatenated), OffTargetSequence (the genomic sequence of the off target), n.mismatch (number of mismatches between the off target and the gRNA), forViewInUCSC (string for viewing in UCSC genome browser, e.g., chr14:31665685-31665707), score (set to 100, and will be updated in getOfftargetScore)

Author(s)

Lihua Julie Zhu

See Also

offTargetAnalysis

Examples

```
all.gRNAs <- findgRNAs(inputFilePath =
    system.file("extdata", "inputseq.fa", package = "CRISPRseek"),
    pairOutputFile = "pairedgRNAs.xls",
```

```
findPairedgRNAOnly = TRUE)

    library("BSgenome.Hsapiens.UCSC.hg19")
    ### for speed reason, use max.mismatch = 0 for finding all targets with
    ### all variants of PAM
    hits <- searchHits(all.gRNAs[1], BSgenomeName = Hsapiens,
        max.mismatch = 0, chromToSearch = "chrX")
    colnames(hits)
```

`translatePattern` *translate pattern from IUPAC Extended Genetic Alphabet to regular expression*

Description

translate pattern containing the IUPAC nucleotide ambiguity codes to regular expression. For example,Y->[C|T], R-> [A|G], S-> [G|C], W-> [A|T], K-> [T|U|G], M-> [A|C], B-> [C|G|T], D-> [A|G|T], H-> [A|C|T], V-> [A|C|G] and N-> [A|C|T|G].

Usage

```
translatePattern(pattern)
```

Arguments

`pattern`	a character vector with the IUPAC nucleotide ambiguity codes

Value

a character vector with the pattern represented as regular expression

Author(s)

Lihua Julie Zhu

Examples

```
pattern1 <- "AACCNWMK"
translatePattern(pattern1)
```

writeHits *Write the hits of sequence search to a file*

Description

write the hits of sequence search to a file, internal function used by searchHits

Usage

```
writeHits(gRNA, seqname, matches, strand, file, gRNA.size = 20,
    PAM = "N[A|G]G$", max.mismatch = 4, chrom.len, append = FALSE)
```

Arguments

gRNA	DNAString object with gRNA sequence with PAM appended immediately after,e.g., ACGTACGTACGTACTGACGTCGG with 20bp gRNA sequence plus 3bp PAM sequence CGG
seqname	chromosome name as character, e.g., chr1
matches	XStringViews object storing matched chromosome locations
strand	strand of the match, + for plus strand and - for minus strand
file	file path where the hits is written to
gRNA.size	gRNA size, default 20
PAM	PAM as regular expression for filtering the hits, default N[A|G]G$
max.mismatch	maximum mismatch allowed within the gRNA (excluding PAM portion) for filtering the hits, default 4
chrom.len	length of the matched chromosome
append	TRUE if append to existing file, false if start a new file

Value

results are saved in the file specified by file

Author(s)

Lihua Julie Zhu

References

http://bioconductor.org/packages/2.8/bioc/vignettes/BSgenome/inst/doc/ GenomeSearching.pdf

See Also

offTargetAnalysis

Examples

```
gRNAPlusPAM <- DNAString("ACGTACGTACGTACTGACGTCGG")
x <- DNAString("AAGCGCGATATGACGTACGTACGTACTGACGTCGG")
chrom.len <- nchar(as.character(x))
m <- matchPattern(gRNAPlusPAM, x)
names(m) <- "testing"
writeHits(gRNA = gRNAPlusPAM, seqname = "chr1",
    matches = m, strand = "+", file = "exampleWriteHits.txt",
    chrom.len = chrom.len, append = FALSE)
```

Index