Package 'GUIDEseq'

November 20, 2024

Type Package

Title GUIDE-seq and PEtag-seq analysis pipeline

Version 1.36.0

Date 2024-04-23

Encoding UTF-8

Author Lihua Julie Zhu, Michael Lawrence, Ankit Gupta, Hervé Pagès , Alper Kucukural, Manuel Garber, Scot A. Wolfe

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

Depends R (>= 3.5.0), GenomicRanges, BiocGenerics

Imports Biostrings, pwalign, CRISPRseek, ChIPpeakAnno, data.table, matrixStats, BSgenome, parallel, IRanges (>= 2.5.5), S4Vectors (>= 0.9.6), stringr, multtest, GenomicAlignments (>= 1.7.3), GenomeInfoDb, Rsamtools, hash, limma,dplyr, GenomicFeatures, rio, tidyr, tools, methods, purrr, ggplot2, openxlsx, patchwork, rlang

biocViews ImmunoOncology, GeneRegulation, Sequencing, WorkflowStep, CRISPR

Suggests knitr, RUnit, BiocStyle, BSgenome.Hsapiens.UCSC.hg19, BSgenome.Hsapiens.UCSC.hg38, TxDb.Hsapiens.UCSC.hg19.knownGene, org.Hs.eg.db, testthat (>= 3.0.0)

VignetteBuilder knitr

Description The package implements GUIDE-seq and PEtag-seq analysis workflow including functions for filtering UMI and reads with low coverage, obtaining unique insertion sites (proxy of cleavage sites), estimating the locations of the insertion sites, aka, peaks, merging estimated insertion sites from plus and minus strand, and performing off target search of the extended regions around insertion sites with mismatches and indels.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation no

Config/testthat/edition 3

RoxygenNote 7.3.1

git_url https://git.bioconductor.org/packages/GUIDEseq

git_branch RELEASE_3_20

git_last_commit fcc97ca

git_last_commit_date 2024-10-29

Repository Bioconductor 3.20

Date/Publication 2024-11-19

Contents

`GUIDEseq-package` *Analysis of GUIDE-seq*

Description

The package includes functions to retain one read per unique molecular identifier (UMI), filter reads lacking integration oligo sequence, identify peak locations (cleavage sites) and heights, merge peaks, perform off-target search using the input gRNA. This package leverages CRISPRseek and ChIPpeakAnno packages.

Details

Package:	GUIDEseq
Type:	Package
Version:	1.0
Date:	2015-09-04
License:	GPL (>= 2)

Function GUIDEseqAnalysis integrates all steps of GUIDE-seq analysis into one function call

Author(s)

Lihua Julie Zhu Maintainer:julie.zhu@umassmed.edu

References

Shengdar Q Tsai and J Keith Joung et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature Biotechnology 33, 187 to 197 (2015)

See Also

GUIDEseqAnalysis

Examples

```
if(interactive())
{
    library("BSgenome.Hsapiens.UCSC.hg19")
    umiFile <- system.file("extdata", "UMI-HEK293_site4_chr13.txt",
       package = "GUIDEseq")
    alignFile <- system.file("extdata","bowtie2.HEK293_site4_chr13.sort.bam" ,
        package = "GUIDEseq")
    gRNA.file <- system.file("extdata","gRNA.fa", package = "GUIDEseq")
    guideSeqRes <- GUIDEseqAnalysis(
        alignment.inputfile = alignFile,
        umi.inputfile = umiFile, gRNA.file = gRNA.file,
        orderOfftargetsBy = "peak_score",
        descending = TRUE,
        keepTopOfftargetsBy = "predicted_cleavage_score",
        scoring.method = "CFDscore",
        BSgenomeName = Hsapiens, min.reads = 80, n.cores.max = 1)
    guideSeqRes$offTargets
}
```

---

annotateOffTargets *Annotate offtargets with gene name*

Description

Annotate offtargets with gene name and whether it is inside an exon

Usage

```
annotateOffTargets(thePeaks, txdb, orgAnn)
```

Arguments

thePeaks	Output from offTargetAnalysisOfPeakRegions
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#___An 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.en for Drosophila and TxDb.Celegans.UCSC.ce6.ensGene for C.elegans

orgAnn — organism annotation mapping such as org.Hs.egSYMBOL in org.Hs.eg.db package for human

Value

A data frame and a tab-delimited file offTargetsInPeakRegions.xls, containing all input offtargets with potential gRNA binding sites, mismatch number and positions, alignment to the input gRNA and predicted cleavage score, and whether the offtargets are inside an exon and associated gene name.

Author(s)

Lihua Julie Zhu

See Also

GUIDEseqAnalysis

Examples

```
if (!interactive()) {
    library("BSgenome.Hsapiens.UCSC.hg19")
    library(TxDb.Hsapiens.UCSC.hg19.knownGene)
    library(org.Hs.eg.db)
    peaks <- system.file("extdata", "T2plus100OffTargets.bed",
        package = "CRISPRseek")
    gRNAs <- system.file("extdata", "T2.fa",
        package = "CRISPRseek")
    outputDir = getwd()
    offTargets <- offTargetAnalysisOfPeakRegions(gRNA = gRNAs, peaks = peaks,
        format=c("fasta", "bed"),
        peaks.withHeader = TRUE, BSgenomeName = Hsapiens,
        upstream = 20L, downstream = 20L, PAM.size = 3L, gRNA.size = 20L,
        orderOfftargetsBy = "predicted_cleavage_score",
        PAM = "NGG", PAM.pattern = "(NGG|NAG|NGA)$", max.mismatch = 2L,
        outputDir = outputDir,
        allowed.mismatch.PAM = 3, overwrite = TRUE)
    annotatedOfftargets <- annotateOffTargets(offTargets,
       txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
       orgAnn = org.Hs.egSYMBOL)
}
```

---

buildFeatureVectorForScoringBulge

Build Feature Vector For Scoring Offtargets with Bulge

Description

Build Feature Vector For Scoring Offtargets with Bulge

Usage

```
buildFeatureVectorForScoringBulge(
  alns,
  gRNA.size = 20,
  canonical.PAM = "NGG",
  subPAM.start = 2,
  subPAM.end = 3,
  insertion.symbol = "^",
  PAM.size = 3,
  PAM.location = "3prime"
)
```

Arguments

alns	alignments, output from getAlnWithBulge (see the example below)
gRNA.size	Size of the gRNA, default to 20L
canonical.PAM	PAM sequence, default to NGG
subPAM.start	start of the subPAM, default to 2L for NGG
subPAM.end	End of the subPAM, default to 3L for NGG
insertion.symbol	Symbol used to indicate bulge in DNA Default to ^
PAM.size	Size of the PAM, default to 3L for NGG
PAM.location	The location of the PAM, default to 3prime

Author(s)

Lihua Julie Zhu

Examples

```
if (interactive())
{
  library(BSgenome.Hsapiens.UCSC.hg19)
  library(GUIDEseq)
  peaks.f <- system.file("extdata", "T2plus100OffTargets.bed",
     package = "GUIDEseq")
  gRNA <- "GACCCCCTCCACCCCGCCTC"
  temp <- GUIDEseq:::getAlnWithBulge(gRNA, gRNA.name = "T2",
      peaks = peaks.f, BSgenomeName = Hsapiens,
       peaks.withHeader = TRUE)
   fv <- buildFeatureVectorForScoringBulge(temp$aln.indel)
   fv$featureVectors
 }
```

combineOfftargets *Combine Offtargets*

Description

Merge offtargets from different samples

Usage

```
combineOfftargets(
  offtarget.folder,
  sample.name,
  remove.common.offtargets = FALSE,
  control.sample.name,
  offtarget.filename = "offTargetsInPeakRegions.xls",
  common.col = c("total.mismatch.bulge", "chromosome", "offTarget_Start",
    "offTarget_End", "offTargetStrand", "offTarget_sequence", "PAM.sequence",
    "guideAlignment2OffTarget", "mismatch.distance2PAM", "n.guide.mismatch",
   "n.PAM.mismatch", "n.DNA.bulge", "n.RNA.bulge", "pos.DNA.bulge", "DNA.bulge",
   "pos.RNA.bulge", "RNA.bulge", "gRNA.name", "gRNAPlusPAM", "predicted_cleavage_score",
    "inExon", "symbol", "entrez_id"),
  exclude.col = "",
  outputFileName,
  comparison.sample1,
  comparison.sample2,
  multiAdjMethod = "BH",
  comparison.score = c("peak_score", "n.distinct.UMIs"),
  overwrite = FALSE
)
```

Arguments

offtarget.folder	offtarget summary output folders created in GUIDEseqAnalysis function
sample.name	Sample names to be used as part of the column names in the final output file
remove.common.offtargets	Default to FALSE If set to TRUE, off-targets common to all samples will be removed.
control.sample.name	The name of the control sample for filtering off-targets present in the control sample
offtarget.filename	Default to offTargetsInPeakRegions.xls, generated in GUIDEseqAnalysis function
common.col	common column names used for merge files. Default to c("total.mismatch.bulge","chromosome", "offTarget_Start","offTarget_End", "offTargetStrand","offTarget_sequence","PAM.sequence","guide "mismatch.distance2PAM","n.guide.mismatch","n.PAM.mismatch", "n.DNA.bulge","n.RNA.bulge", "RNA.bulge","gRNA.name","gRNAPlusPAM","predicted_cleavage_score", "in-Exon","symbol","entrez_id")

`exclude.col` columns to be excluded before merging. Please check offTargetsInPeakRegions.xls to choose the desired columns to exclude

`outputFileName` The merged offtarget file

`comparison.sample1`
A vector of sample names to be used for comparison. For example, comparison.sample1 = c("A", "B"), comparison.sample2 = rep("Control", 2) indicates that you are interested in comparing sample A vs Control and B vs Control Please make sure the sample names specified in comparison.sample1 and comparison.sample2 are in the sample name list specified in sample.name

`comparison.sample2`
A vector of sample names to be used for comparison. For example, comparison.sample1 = c("A", "B"), comparison.sample2 = rep("Control", 2) indicates that you are interested in comparing sample A vs Control and B vs Control

`multiAdjMethod` A vector of character strings containing the names of the multiple testing procedures for which adjusted p-values are to be computed. This vector should include any of the following: "none", "Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD", "BH", "BY", "ABH", and "TSBH". Please type ?multtest::mt.rawp2adjp for details. Default to "BH"

`comparison.score`
the score to be used for statistical analysis. Two options are available: "peak_score" and "n.distinct.UMIs" n.distinct.UMIs is the number of unique UMIs in the associated peak region without considering the sequence coordinates while peak_score takes into consideration of the sequence coordinates

`overwrite` Indicates whether to overwrite the existing file specified by outputFileName, default to FALSE.

Details

Please note that by default, merged file will only contain peaks with offtargets found in the genome in GUIDEseqAnalysis function.

Value

a data frame containing all off-targets from all samples merged by the columns specified in common.col. Sample specific columns have sample.name concatenated to the original column name, e.g., peak_score becomes sample1.peak_score.

Author(s)

Lihua Julie Zhu

Examples

```
  offtarget.folder <- system.file("extdata",
      c("sample1-17", "sample2-18", "sample3-19"),
      package = "GUIDEseq")
  mergedOfftargets <-
     combineOfftargets(offtarget.folder = offtarget.folder,
     sample.name = c("Cas9Only", "WT-SpCas9", "SpCas9-MT3-ZFP"),
comparison.sample1 = c("Cas9Only", "SpCas9-MT3-ZFP"),
comparison.sample2 = rep("WT-SpCas9", 2),
     outputFileName = "TS2offtargets3Constructs.xlsx")
```

compareSamples *Compare Samples using Fisher's exact test*

Description

Compare Samples using Fisher's exact test

Usage

```
compareSamples(
  df,
  col.count1,
  col.count2,
  total1,
  total2,
  multiAdjMethod = "BH",
  comparison.score = c("peak_score", "umi.count")
)
```

Arguments

df	a data frame containing the peak score and sequence depth for each sample
col.count1	the score (e.g., peak_score) column used as the numerator for calculating odds ratio. For example,if the tenth column contains the score for sample 1, then set col.count1 = 10
col.count2	the score (e.g., peak_score) column used as the denominator for calculating odds ratio. For example,if the nineteenth column contains the score for sample 1, then set col.count2 = 19
total1	the sequence depth for sample 1
total2	the sequence depth for sample 2
multiAdjMethod	A vector of character strings containing the names of the multiple testing procedures for which adjusted p-values are to be computed. This vector should include any of the following: "none", "Bonferroni", "Holm", "Hochberg", "SidakSS", "SidakSD", "BH", "BY", "ABH", and "TSBH". Please type ?multtest::mt.rawp2adjp for details. Default to "BH"
comparison.score	the score to be used for statistical analysis. Two options are available: "peak_score" and "umi.count" umi.count is the number of unique UMIs in the associated peak region without considering the sequence coordinates while peak_score takes into consideration of the sequence coordinates

Author(s)

Lihua Julie Zhu

createBarcodeFasta *Create barcode as fasta file format for building bowtie1 index*

Description

Create barcode as fasta file format for building bowtie1 index to assign reads to each library with different barcodes. The bowtie1 index has been built for the standard GUIDE-seq protocol using the standard p5 and p7 index. It can be downloaded at http://mccb.umassmed.edu/GUIDE-seq/barcode.bowtie1.index.tar.gz

Usage

```
createBarcodeFasta(
  p5.index,
  p7.index,
  reverse.p7 = TRUE,
  reverse.p5 = FALSE,
  header = FALSE,
  outputFile = "barcodes.fa"
)
```

Arguments

p5.index	A text file with one p5 index sequence per line
p7.index	A text file with one p7 index sequence per line
reverse.p7	Indicate whether to reverse p7 index, default to TRUE for standard GUIDE-seq experiments
reverse.p5	Indicate whether to reverse p5 index, default to FALSE for standard GUIDE-seq experiments
header	Indicate whether there is a header in the p5.index and p7.index files. Default to FALSE
outputFile	Give a name to the output file where the generated barcodes are written. This file can be used to build bowtie1 index for binning reads.

Note

Create barcode file to be used to bin the reads sequenced in a pooled lane

Author(s)

Lihua Julie Zhu

Examples

```
p7 <- system.file("extdata", "p7.index",
       package = "GUIDEseq")
p5 <- system.file("extdata", "p5.index",
       package = "GUIDEseq")
outputFile <- "barcodes.fa"
createBarcodeFasta(p5.index = p5, p7.index = p7, reverse.p7 = TRUE,
    reverse.p5 = FALSE, outputFile = outputFile)
```

`getBestAlnInfo` *Parse pairwise alignment*

Description

Parse pairwise alignment

Usage

```
getBestAlnInfo(
  offtargetSeq,
  pa.f,
  pa.r,
  gRNA.size = 20,
  PAM = "NGG",
  PAM.size = 3,
  insertion.symbol = "^"
)
```

Arguments

`offtargetSeq`	DNAStringSet object of length 1
`pa.f`	Global-Local PairwiseAlignmentsSingleSubject, results of pairwiseAlignment, alignment of pattern to subject
`pa.r`	Global-Local PairwiseAlignmentsSingleSubject, results of pairwiseAlignment, alignment of pattern to reverse subject
`gRNA.size`	size of gRNA, default to 20
`PAM`	PAM sequence, default to NGG
`PAM.size`	PAM size, default to 3
`insertion.symbol`	symbol for representing bulge in offtarget, default to ^. It can also be set to lowerCase to use lower case letter to represent insertion

Value

a dataframe with the following columns. offTarget: name of the offtarget peak_score: place holder for storing peak score gRNA.name: place holder for storing gRNA name gRNAPlusPAM: place holder for storing gRNAPlusPAM sequence offTarget_sequence: offTarget sequence with PAM in the right orientation. For PAM in the 3' prime location, offTarget is the sequence on the plus strand otherwise, is the sequence on the reverse strand seq.aligned: the aligned sequence without PAM guideAlignment2OffTarget: string representation of the alignment offTargetStrand: the strand of the offtarget mismatch.distance2PAM: mismatch distance to PAM start n.PAM.mismatch: number of mismatches in PAM n.guide.mismatch: number of mismatches in the gRNA not including PAM PAM.sequence: PAM in the offtarget offTarget_Start: offtarget start offTarget_End: offTarget end chromosome: place holder for storing offtarget chromosome pos.mismatch: mismatch positions with the correct PAM orientation, i.e., indexed form distal to proximal of PAM pos.indel: indel positions starting with deletions in the gRNA followed by those in the offtarget pos.insertion: Insertion positions in the gRNA Insertion positions are counted from distal to proximal of PAM For example, 5 means the 5th position is an insertion in gRNA pos.deletion: Deletion in the gRNA Deletion

positions are counted from distal to proximal of PAM For example, 5 means the 5th position is a deletion in gRNA n.insertion: Number of insertions in the RNA. Insertions in gRNA creates bulged DNA base n.deletion: Number of deletions in the RNA. Deletions in gRNA creates bulged DNA base

Author(s)

Lihua Julie Zhu

`getPeaks` *Obtain peaks from GUIDE-seq*

Description

Obtain strand-specific peaks from GUIDE-seq

Usage

```
getPeaks(
  gr,
  window.size = 20L,
  step = 20L,
  bg.window.size = 5000L,
  min.reads = 10L,
  min.SNratio = 2,
  maxP = 0.05,
  stats = c("poisson", "nbinom"),
 p.adjust.methods = c("none", "BH", "holm", "hochberg", "hommel", "bonferroni", "BY",
    "fdr")
)
```

Arguments

`gr`	GRanges with cleavage sites, output from getUniqueCleavageEvents
`window.size`	window size to calculate coverage
`step`	step size to calculate coverage
`bg.window.size`	window size to calculate local background
`min.reads`	minimum number of reads to be considered as a peak
`min.SNratio`	minimum signal noise ratio, which is the coverage normalized by local background
`maxP`	Maximum p-value to be considered as significant
`stats`	Statistical test, default poisson
`p.adjust.methods`	Adjustment method for multiple comparisons, default none

Value

peaks — GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

summarized.count — A data frame contains the same information as peaks except that it has all the sites without filtering.

Author(s)

Lihua Julie Zhu

Examples

```
if (interactive())
{
    data(uniqueCleavageEvents)
    peaks <- getPeaks(uniqueCleavageEvents$cleavage.gr,
        min.reads = 80)
    peaks$peaks
}
```

---

getUniqueCleavageEvents

Using UMI sequence to obtain the starting sequence library

Description

PCR amplification often leads to biased representation of the starting sequence population. To track the sequence tags present in the initial sequence library, a unique molecular identifier (UMI) is added to the 5 prime of each sequence in the staring library. This function uses the UMI sequence plus the first few sequence from R1 reads to obtain the starting sequence library.

Usage

```
getUniqueCleavageEvents(
  alignment.inputfile,
  umi.inputfile,
  alignment.format = c("auto", "bam", "bed"),
  umi.header = FALSE,
  read.ID.col = 1,
  umi.col = 2,
  umi.sep = "\t",
  keep.chrM = FALSE,
  keep.R1only = TRUE,
  keep.R2only = TRUE,
  concordant.strand = TRUE,
  max.paired.distance = 1000,
  min.mapping.quality = 30,
  max.R1.len = 130,
  max.R2.len = 130,
```

```
  apply.both.max.len = FALSE,
  same.chromosome = TRUE,
  distance.inter.chrom = -1,
  min.R1.mapped = 20,
  min.R2.mapped = 20,
  apply.both.min.mapped = FALSE,
  max.duplicate.distance = 0L,
  umi.plus.R1start.unique = TRUE,
  umi.plus.R2start.unique = TRUE,
  min.umi.count = 5L,
  max.umi.count = 100000L,
  min.read.coverage = 1L,
  n.cores.max = 6,
  outputDir,
  removeDuplicate = TRUE,
  ignoreTagmSite = FALSE,
  ignoreUMI = FALSE
)
```

Arguments

`alignment.inputfile`
: The alignment file. Currently supports bed output file with CIGAR information. Suggest run the workflow binReads.sh, which sequentially runs barcode binning, adaptor removal, alignment to genome, alignment quality filtering, and bed file conversion. Please download the workflow function and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/binReads/

`umi.inputfile`
: A text file containing at least two columns, one is the read identifier and the other is the UMI or UMI plus the first few bases of R1 reads. Suggest use getUMI.sh to generate this file. Please download the script and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/getUMI/

`alignment.format`
: The format of the alignment input file. Currently only bam and bed file format is supported. BED format will be deprecated soon.

`umi.header`
: Indicates whether the umi input file contains a header line or not. Default to FALSE

`read.ID.col`
: The index of the column containing the read identifier in the umi input file, default to 1

`umi.col`
: The index of the column containing the umi or umi plus the first few bases of sequence from the R1 reads, default to 2

`umi.sep`
: column separator in the umi input file, default to tab

`keep.chrM`
: Specify whether to include alignment from chrM. Default FALSE

`keep.R1only`
: Specify whether to include alignment with only R1 without paired R2. Default TRUE

`keep.R2only`
: Specify whether to include alignment with only R2 without paired R1. Default TRUE

`concordant.strand`
: Specify whether the R1 and R2 should be aligned to the same strand or opposite strand. Default opposite.strand (TRUE)

`max.paired.distance`
: Specify the maximum distance allowed between paired R1 and R2 reads. Default 1000 bp

`min.mapping.quality`
: Specify min.mapping.quality of acceptable alignments

`max.R1.len`
: The maximum retained R1 length to be considered for downstream analysis, default 130 bp. Please note that default of 130 works well when the read length 150 bp. Please also note that retained R1 length is not necessarily equal to the mapped R1 length

`max.R2.len`
: The maximum retained R2 length to be considered for downstream analysis, default 130 bp. Please note that default of 130 works well when the read length 150 bp. Please also note that retained R2 length is not necessarily equal to the mapped R2 length

`apply.both.max.len`
: Specify whether to apply maximum length requirement to both R1 and R2 reads, default FALSE

`same.chromosome`
: Specify whether the paired reads are required to align to the same chromosome, default TRUE

`distance.inter.chrom`
: Specify the distance value to assign to the paired reads that are aligned to different chromosome, default -1

`min.R1.mapped`
: The maximum mapped R1 length to be considered for downstream analysis, default 30 bp.

`min.R2.mapped`
: The maximum mapped R2 length to be considered for downstream analysis, default 30 bp.

`apply.both.min.mapped`
: Specify whether to apply minimum mapped length requirement to both R1 and R2 reads, default FALSE

`max.duplicate.distance`
: Specify the maximum distance apart for two reads to be considered as duplicates, default 0. Currently only 0 is supported

`umi.plus.R1start.unique`
: To specify whether two mapped reads are considered as unique if both containing the same UMI and same alignment start for R1 read, default TRUE.

`umi.plus.R2start.unique`
: To specify whether two mapped reads are considered as unique if both containing the same UMI and same alignment start for R2 read, default TRUE.

`min.umi.count`
: To specify the minimum count for a umi to be included in the subsequent analysis. Please adjust it to a higher number for deeply sequenced library and vice versa.

`max.umi.count`
: To specify the maximum count for a umi to be included in the subsequent analysis. Please adjust it to a higher number for deeply sequenced library and vice versa.

`min.read.coverage`
: To specify the minimum coverage for a read UMI combination to be included in the subsequent analysis. Please note that this is different from min.umi.count which is less stringent.

`n.cores.max` Indicating maximum number of cores to use in multi core mode, i.e., parallel processing, default 6. Please set it to 1 to disable multicore processing for small dataset.

`outputDir` output Directory to save the figures

`removeDuplicate` default to TRUE. Set it to FALSE if PCR duplicates not to be removed for testing purpose.

`ignoreTagmSite` default to FALSE. To collapse reads with the same integration site and UMI but with different tagmentation site, set the option to TRUE.

`ignoreUMI` default to FALSE. To collapse reads with the same integration and tagmentation site but with different UMIs, set the option to TRUE and retain the UMI that appears most frequently for each combination of integration and tagmentation site. In case of ties, randomly select one UMI.

Value

`cleavage.gr` Cleavage sites with one site per UMI as GRanges with metadata column total set to 1 for each range

`unique.umi.plus.R2` a data frame containing unique cleavage site from R2 reads mapped to plus strand with the following columns: seqnames (chromosome), start (cleavage/Integration site), strand, UMI (unique molecular identifier), and UMI read duplication level (min.read.coverage can be used to remove UMI-read with very low coverage)

`unique.umi.minus.R2` a data frame containing unique cleavage site from R2 reads mapped to minus strand with the same columns as unique.umi.plus.R2

`unique.umi.plus.R1` a data frame containing unique cleavage site from R1 reads mapped to minus strand without corresponding R2 reads mapped to the plus strand, with the same columns as unique.umi.plus.R2

`unique.umi.minus.R1` a data frame containing unique cleavage site from R1 reads mapped to plus strand without corresponding R2 reads mapped to the minus strand, with the same columns as unique.umi.plus.R2

`align.umi` a data frame containing all the mapped reads with the following columns. readName (read ID), chr.x (chromosome of readSide.x/R1 read), start.x (start of eadSide.x/R1 read), end.x (end of eadSide.x/R1 read), mapping.qual.x (mapping quality of readSide.x/R1 read), strand.x (strand of readSide.x/R1 read), cigar.x (CIGAR of readSide.x/R1 read) , readSide.x (1/R1), chr.y (chromosome of readSide.y/R2 read) start.y (start of readSide.y/R2 read), end.y (end of readSide.y/R2 read), mapping.qual.y (mapping quality of readSide.y/R2 read), strand.y (strand of readSide.y/R2 read), cigar.y (CIGAR of readSide.y/R2 read), readSide.y (2/R2) R1.base.kept (retained R1 length), R2.base.kept (retained R2 length), distance (distance between mapped R1 and R2), UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

Author(s)

Lihua Julie Zhu

References

Shengdar Q Tsai and J Keith Joung et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature Biotechnology 33, 187 to 197 (2015)

See Also

getPeaks

Examples

```
if(interactive())
{
    umiFile <- system.file("extdata", "UMI-HEK293_site4_chr13.txt",
       package = "GUIDEseq")
    alignFile <- system.file("extdata","bowtie2.HEK293_site4_chr13.sort.bam" ,
        package = "GUIDEseq")
    cleavages <- getUniqueCleavageEvents(
        alignment.inputfile = alignFile , umi.inputfile = umiFile,
        n.cores.max = 1)
    names(cleavages)
    #output a summary of duplicate counts for sequencing saturation assessment
    table(cleavages$umi.count.summary$n)
}
```

`getUsedBarcodes` *Create barcodes from the p5 and p7 index used for each sequencing lane*

Description

Create barcodes from the p5 and p7 index for assigning reads to each barcode

Usage

```
getUsedBarcodes(
  p5.index,
  p7.index,
  header = FALSE,
  reverse.p7 = TRUE,
  reverse.p5 = FALSE,
  outputFile
)
```

Arguments

`p5.index`	A text file with one p5 index sequence per line
`p7.index`	A text file with one p7 index sequence per line
`header`	Indicate whether there is a header in the p5.index and p7.index files. Default to FALSE
`reverse.p7`	Indicate whether to reverse p7 index, default to TRUE for standard GUIDE-seq experiments

reverse.p5	Indicate whether to reverse p5 index, default to FALSE for standard GUIDE-seq experiments
outputFile	Give a name to the output file where the generated barcodes are written

Value

DNAStringSet

Note

Create barcode file to be used to bin the reads sequenced in a pooled lane

Author(s)

Lihua Julie Zhu

Examples

```
p7 <- system.file("extdata", "p7.index",
       package = "GUIDEseq")
p5 <- system.file("extdata", "p5.index",
       package = "GUIDEseq")
outputFile <- "usedBarcode"
getUsedBarcodes(p5.index = p5, p7.index = p7, reverse.p7 = TRUE,
    reverse.p5 = FALSE, outputFile = outputFile)
```

GUIDEseqAnalysis *Analysis pipeline for GUIDE-seq dataset*

Description

A wrapper function that uses the UMI sequence plus the first few bases of each sequence from R1 reads to estimate the starting sequence library, piles up reads with a user defined window and step size, identify the insertion sites (proxy of cleavage sites), merge insertion sites from plus strand and minus strand, followed by off target analysis of extended regions around the identified insertion sites.

Usage

```
GUIDEseqAnalysis(
  alignment.inputfile,
  umi.inputfile,
  alignment.format = c("auto", "bam", "bed"),
  umi.header = FALSE,
  read.ID.col = 1L,
  umi.col = 2L,
  umi.sep = "\t",
  BSgenomeName,
  gRNA.file,
  outputDir,
  n.cores.max = 1L,
  keep.chrM = FALSE,
```

```
 keep.R1only = TRUE,
 keep.R2only = TRUE,
 concordant.strand = TRUE,
 max.paired.distance = 1000L,
 min.mapping.quality = 30L,
 max.R1.len = 130L,
 max.R2.len = 130L,
 min.umi.count = 1L,
 max.umi.count = 100000L,
 min.read.coverage = 1L,
 apply.both.max.len = FALSE,
 same.chromosome = TRUE,
 distance.inter.chrom = -1L,
 min.R1.mapped = 20L,
 min.R2.mapped = 20L,
 apply.both.min.mapped = FALSE,
 max.duplicate.distance = 0L,
 umi.plus.R1start.unique = TRUE,
 umi.plus.R2start.unique = TRUE,
 window.size = 20L,
 step = 20L,
 bg.window.size = 5000L,
 min.reads = 5L,
 min.reads.per.lib = 1L,
 min.peak.score.1strandOnly = 5L,
 min.SNratio = 2,
 maxP = 0.01,
 stats = c("poisson", "nbinom"),
p.adjust.methods = c("none", "BH", "holm", "hochberg", "hommel", "bonferroni", "BY",
   "fdr"),
 distance.threshold = 40L,
 max.overlap.plusSig.minusSig = 30L,
 plus.strand.start.gt.minus.strand.end = TRUE,
 keepPeaksInBothStrandsOnly = TRUE,
 gRNA.format = "fasta",
 overlap.gRNA.positions = c(17, 18),
 upstream = 25L,
 downstream = 25L,
 PAM.size = 3L,
 gRNA.size = 20L,
 PAM = "NGG",
 PAM.pattern = "NNN$",
 max.mismatch = 6L,
 allowed.mismatch.PAM = 2L,
 overwrite = TRUE,
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),
orderOfftargetsBy = c("peak_score", "predicted_cleavage_score", "n.guide.mismatch"),
 descending = TRUE,
 keepTopOfftargetsOnly = TRUE,
 keepTopOfftargetsBy = c("predicted_cleavage_score", "n.mismatch"),
 scoring.method = c("Hsu-Zhang", "CFDscore"),
```

```
 subPAM.activity = hash(AA = 0, AC = 0, AG = 0.259259259, AT = 0, CA = 0, CC = 0, CG =
  0.107142857, CT = 0, GA = 0.069444444, GC = 0.022222222, GG = 1, GT = 0.016129032, TA
    = 0, TC = 0, TG = 0.038961039, TT = 0),
 subPAM.position = c(22, 23),
 PAM.location = "3prime",
 mismatch.activity.file = system.file("extdata",
   "NatureBiot2016SuppTable19DoenchRoot.csv", package = "CRISPRseek"),
 bulge.activity.file = system.file("extdata",
   "NatureBiot2016SuppTable19DoenchRoot.xlsx", package = "GUIDEseq"),
 txdb,
 orgAnn,
 mat,
 includeBulge = FALSE,
 max.n.bulge = 2L,
 min.peak.score.bulge = 60L,
 removeDuplicate = TRUE,
 resume = FALSE,
 ignoreTagmSite = FALSE,
 ignoreUMI = FALSE
)
```

Arguments

`alignment.inputfile`
: The alignment file. Currently supports bam and bed output file with CIGAR information. Suggest run the workflow binReads.sh, which sequentially runs barcode binning, adaptor removal, alignment to genome, alignment quality filtering, and bed file conversion. Please download the workflow function and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/binReads/

`umi.inputfile`
: A text file containing at least two columns, one is the read identifier and the other is the UMI or UMI plus the first few bases of R1 reads. Suggest use getUMI.sh to generate this file. Please download the script and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/getUMI/

`alignment.format`
: The format of the alignment input file. Default bed file format. Currently only bed file format is supported, which is generated from binReads.sh

`umi.header`
: Indicates whether the umi input file contains a header line or not. Default to FALSE

`read.ID.col`
: The index of the column containing the read identifier in the umi input file, default to 1

`umi.col`
: The index of the column containing the umi or umi plus the first few bases of sequence from the R1 reads, default to 2

`umi.sep`
: column separator in the umi input file, default to tab

`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

`gRNA.file`
: gRNA input file path or a DNAStringSet object that contains the target sequence (gRNA plus PAM)

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

`n.cores.max` Indicating maximum number of cores to use in multi core mode, i.e., parallel processing, default 1 to disable multicore processing for small dataset.

`keep.chrM` Specify whether to include alignment from chrM. Default FALSE

`keep.R1only` Specify whether to include alignment with only R1 without paired R2. Default TRUE

`keep.R2only` Specify whether to include alignment with only R2 without paired R1. Default TRUE

`concordant.strand`
Specify whether the R1 and R2 should be aligned to the same strand or opposite strand. Default opposite.strand (TRUE)

`max.paired.distance`
Specify the maximum distance allowed between paired R1 and R2 reads. Default 1000 bp

`min.mapping.quality`
Specify min.mapping.quality of acceptable alignments

`max.R1.len` The maximum retained R1 length to be considered for downstream analysis, default 130 bp. Please note that default of 130 works well when the read length 150 bp. Please also note that retained R1 length is not necessarily equal to the mapped R1 length

`max.R2.len` The maximum retained R2 length to be considered for downstream analysis, default 130 bp. Please note that default of 130 works well when the read length 150 bp. Please also note that retained R2 length is not necessarily equal to the mapped R2 length

`min.umi.count` To specify the minimum total count for a umi at the genome level to be included in the subsequent analysis. For example, with min.umi.count set to 2, if a umi only has 1 read in the entire genome, then that umi will be excluded for the subsequent analysis. Please adjust it to a higher number for deeply sequenced library and vice versa.

`max.umi.count` To specify the maximum count for a umi to be included in the subsequent analysis. Please adjust it to a higher number for deeply sequenced library and vice versa.

`min.read.coverage`
To specify the minimum coverage for a read UMI combination to be included in the subsequent analysis. Please note that this is different from min.umi.count which is less stringent.

`apply.both.max.len`
Specify whether to apply maximum length requirement to both R1 and R2 reads, default FALSE

`same.chromosome`
Specify whether the paired reads are required to align to the same chromosome, default TRUE

`distance.inter.chrom`
Specify the distance value to assign to the paired reads that are aligned to different chromosome, default -1

`min.R1.mapped` The minimum mapped R1 length to be considered for downstream analysis, default 30 bp.

`min.R2.mapped` The minimum mapped R2 length to be considered for downstream analysis, default 30 bp.

apply.both.min.mapped
: Specify whether to apply minimum mapped length requirement to both R1 and R2 reads, default FALSE

max.duplicate.distance
: Specify the maximum distance apart for two reads to be considered as duplicates, default 0. Currently only 0 is supported

umi.plus.R1start.unique
: To specify whether two mapped reads are considered as unique if both containing the same UMI and same alignment start for R1 read, default TRUE.

umi.plus.R2start.unique
: To specify whether two mapped reads are considered as unique if both containing the same UMI and same alignment start for R2 read, default TRUE.

window.size
: window size to calculate coverage

step
: step size to calculate coverage

bg.window.size
: window size to calculate local background

min.reads
: minimum number of reads to be considered as a peak

min.reads.per.lib
: minimum number of reads in each library (usually two libraries) to be considered as a peak

min.peak.score.1strandOnly
: Specify the minimum number of reads for a one-strand only peak to be included in the output. Applicable when set keepPeaksInBothStrandsOnly to FALSE and there is only one library per sample

min.SNratio
: Specify the minimum signal noise ratio to be called as peaks, which is the coverage normalized by local background.

maxP
: Specify the maximum p-value to be considered as significant

stats
: Statistical test, currently only poisson is implemented

p.adjust.methods
: Adjustment method for multiple comparisons, default none

distance.threshold
: Specify the maximum gap allowed between the plus strand and the negative strand peak, default 40. Suggest set it to twice of window.size used for peak calling.

max.overlap.plusSig.minusSig
: Specify the cushion distance to allow sequence error and inprecise integration Default to 30 to allow at most 10 (30-window.size 20) bp (half window) of minus-strand peaks on the right side of plus-strand peaks. Only applicable if plus.strand.start.gt.minus.strand.end is set to TRUE.

plus.strand.start.gt.minus.strand.end
: Specify whether plus strand peak start greater than the paired negative strand peak end. Default to TRUE

keepPeaksInBothStrandsOnly
: Indicate whether only keep peaks present in both strands as specified by plus.strand.start.gt.minus.stra
max.overlap.plusSig.minusSig and distance.threshold.

gRNA.format
: Format of the gRNA input file. Currently, fasta is supported

overlap.gRNA.positions
: The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18 for SpCas9.

`upstream` upstream offset from the peak start to search for off targets, default 25 suggest set it to window size

`downstream` downstream offset from the peak end to search for off targets, default 25 suggest set it to window size

`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 protospacer-adjacent motif (PAM), default NNN$. Alternatively set it to (NAG|NGG|NGA)$ for off target search

`max.mismatch` Maximum mismatch to the gRNA (not including mismatch to the PAM) allowed in off target search, default 6

`allowed.mismatch.PAM`
Maximum number of mismatches allowed for the PAM sequence plus the number of degenerate sequence in the PAM sequence, default to 2 for NGG PAM

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

`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) for SPcas9 system, which is used in Hsu et al., 2013 cited in the reference section. Please make sure that the number of elements in this vector is the same as the gRNA.size, e.g., pad 0s at the beginning of the vector.

`orderOfftargetsBy`
Criteria to order the offtargets, which works together with the descending parameter

`descending` Indicate the output order of the offtargets, i.e., in the descending or ascending order.

`keepTopOfftargetsOnly`
Output all offtargets or the top offtarget using the keepOfftargetsBy criteria, default to the top offtarget

`keepTopOfftargetsBy`
Output the top offtarget for each called peak using the keepTopOfftargetsBy criteria, If set to predicted_cleavage_score, then the offtargets with the highest predicted cleavage score will be retained If set to n.mismatch, then the offtarget with the lowest number of mismatch to the target sequence will be retained

`scoring.method` Indicates which method to use for offtarget cleavage rate estimation, currently two methods are supported, Hsu-Zhang and CFDscore

`subPAM.activity`
Applicable only when scoring.method is set to CFDscore A hash to represent the cleavage rate for each alternative sub PAM sequence relative to preferred PAM sequence

`subPAM.position`
Applicable only when scoring.method is set to CFDscore The start and end positions of the sub PAM. Default to 22 and 23 for SP with 20bp gRNA and NGG as preferred PAM

`PAM.location` PAM location relative to gRNA. For example, default to 3prime for spCas9 PAM. Please set to 5prime for cpf1 PAM since it's PAM is located on the 5 prime end

`mismatch.activity.file`
: Applicable only when scoring.method is set to CFDscore A comma separated (csv) file containing the cleavage rates for all possible types of single nucleotide mismatche at each position of the gRNA. By default, use the supplemental Table 19 from Doench et al., Nature Biotechnology 2016

`bulge.activity.file`
: Used for predicting indel effect on offtarget cleavage score. An excel file with the second sheet for deletion activity and the third sheet for Insertion. By default, use the supplemental Table 19 from Doench et al., Nature Biotechnology 2016

`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#___An 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.en for Drosophila and TxDb.Celegans.UCSC.ce6.ensGene for C.elegans

`orgAnn`
: organism annotation mapping such as org.Hs.egSYMBOL in org.Hs.eg.db package for human

`mat`
: nucleotide substitution matrix. Function nucleotideSubstitutionMatrix can be used for creating customized nucleotide substitution matrix. By default, match = 1, mismatch = -1, and baseOnly = TRUE Only applicalbe with includeBulge set to TRUE

`includeBulge`
: indicates whether including offtargets with indels default to FALSE

`max.n.bulge`
: offtargets with at most this number of indels to be included in the offtarget list. Only applicalbe with includeBulge set to TRUE

`min.peak.score.bulge`
: default to 60. Set it to a higher number to speed up the alignment with bulges. Any peaks with peak.score less than min.peak.score.bulge will not be included in the offtarget analysis with bulges. However, all peaks will be included in the offtarget analysis with mismatches.

`removeDuplicate`
: default to TRUE. Set it to FALSE if PCR duplicates not to be removed for testing purpose

`resume`
: default to FALSE to restart the analysis. set it TRUE to resume an analysis.

`ignoreTagmSite`
: default to FALSE. To collapse reads with the same integration site and UMI but with different tagmentation site, set the option to TRUE.

`ignoreUMI`
: default to FALSE. To collapse reads with the same integration and tagmentation site but with different UMIs, set the option to TRUE and retain the UMI that appears most frequently for each combination of integration and tagmentation site. In case of ties, randomly select one UMI.

Value

`offTargets`
: a data frame, containing all input peaks with potential gRNA binding sites, mismatch number and positions, alignment to the input gRNA and predicted cleavage score.

`merged.peaks`
: merged peaks as GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

`peaks`
: GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

`uniqueCleavages` Cleavage sites with one site per UMI as GRanges with metadata column total set to 1 for each range

`read.summary` One table per input mapping file that contains the number of reads for each chromosome location

`sequence.depth` sequence depth in the input alignment files

Author(s)

Lihua Julie Zhu

References

Lihua Julie Zhu, Michael Lawrence, Ankit Gupta, Herve Pages, Alper Ku- cukural, Manuel Garber and Scot A. Wolfe. GUIDEseq: a bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases. BMC Genomics. 2017. 18:379

See Also

getPeaks

Examples

```
if(interactive())
    {
        library("BSgenome.Hsapiens.UCSC.hg19")
        umiFile <- system.file("extdata", "UMI-HEK293_site4_chr13.txt",
           package = "GUIDEseq")
        alignFile <- system.file("extdata","bowtie2.HEK293_site4_chr13.sort.bam" ,
            package = "GUIDEseq")
        gRNA.file <- system.file("extdata","gRNA.fa", package = "GUIDEseq")
        guideSeqRes <- GUIDEseqAnalysis(
            alignment.inputfile = alignFile,
            umi.inputfile = umiFile, gRNA.file = gRNA.file,
            orderOfftargetsBy = "peak_score",
            descending = TRUE,
            keepTopOfftargetsBy = "predicted_cleavage_score",
            scoring.method = "CFDscore",
            BSgenomeName = Hsapiens, min.reads = 80, n.cores.max = 1)
        guideSeqRes$offTargets
        names(guideSeqRes)
   }
```

---

mergePlusMinusPeaks *Merge peaks from plus strand and minus strand*

Description

Merge peaks from plus strand and minus strand with required orientation and within certain distance apart

Usage

```
mergePlusMinusPeaks(
  peaks.gr,
  peak.height.mcol = "count",
  bg.height.mcol = "bg",
  distance.threshold = 40L,
  max.overlap.plusSig.minusSig = 30L,
  plus.strand.start.gt.minus.strand.end = TRUE,
  output.bedfile
)
```

Arguments

peaks.gr	Specify the peaks as GRanges object, which should contain peaks from both plus and minus strand. In addition, it should contain peak height metadata column to store peak height and optionally background height.
peak.height.mcol	Specify the metadata column containing the peak height, default to count
bg.height.mcol	Specify the metadata column containing the background height, default to bg
distance.threshold	Specify the maximum gap allowed between the plus stranded and the negative stranded peak, default 40. Suggest set it to twice of window.size used for peak calling.
max.overlap.plusSig.minusSig	Specify the cushion distance to allow sequence error and inprecise integration Default to 30 to allow at most 10 (30-window.size 20) bp (half window) of minus-strand peaks on the right side of plus-strand peaks. Only applicable if plus.strand.start.gt.minus.strand.end is set to TRUE.
plus.strand.start.gt.minus.strand.end	Specify whether plus strand peak start greater than the paired negative strand peak end. Default to TRUE
output.bedfile	Specify the bed output file name, which is used for off target analysis subsequently.

Value

output a list and a bed file containing the merged peaks a data frame of the bed format

mergedPeaks.gr	merged peaks as GRanges
mergedPeaks.bed	merged peaks in bed format

Author(s)

Lihua Julie Zhu

References

Zhu L.J. et al. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 2010, 11:237doi:10.1186/1471-2105-11-237. Zhu L.J. (2013) Integrative analysis of ChIP-chip and ChIP-seq dataset. Methods Mol Biol. 2013;1067:105-24. doi: 10.1007/978-1-62703-607-8_8.

Examples

```
if (interactive())
{
    data(peaks.gr)
    mergedPeaks <- mergePlusMinusPeaks(peaks.gr = peaks.gr,
        output.bedfile = "mergedPeaks.bed")
    mergedPeaks$mergedPeaks.gr
    head(mergedPeaks$mergedPeaks.bed)
}
```

---

offTargetAnalysisOfPeakRegions

Offtarget Analysis of GUIDE-seq peaks

Description

Finding offtargets around peaks from GUIDE-seq or around any given genomic regions

Usage

```
offTargetAnalysisOfPeakRegions(
  gRNA,
  peaks,
  format = c("fasta", "bed"),
  peaks.withHeader = FALSE,
  BSgenomeName,
  overlap.gRNA.positions = c(17, 18),
  upstream = 25L,
  downstream = 25L,
  PAM.size = 3L,
  gRNA.size = 20L,
  PAM = "NGG",
  PAM.pattern = "NNN$",
  max.mismatch = 6L,
  outputDir,
  allowed.mismatch.PAM = 2L,
  overwrite = TRUE,
 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),
  orderOfftargetsBy = c("predicted_cleavage_score", "n.mismatch"),
  descending = TRUE,
  keepTopOfftargetsOnly = TRUE,
  scoring.method = c("Hsu-Zhang", "CFDscore"),
 subPAM.activity = hash(AA = 0, AC = 0, AG = 0.259259259, AT = 0, CA = 0, CC = 0, CG =
  0.107142857, CT = 0, GA = 0.069444444, GC = 0.022222222, GG = 1, GT = 0.016129032, TA
    = 0, TC = 0, TG = 0.038961039, TT = 0),
  subPAM.position = c(22, 23),
  PAM.location = "3prime",
  mismatch.activity.file = system.file("extdata",
```

```
    "NatureBiot2016SuppTable19DoenchRoot.csv", package = "CRISPRseek"),
  n.cores.max = 1
)
```

Arguments

gRNA	gRNA input file path or a DNAStringSet object that contains gRNA plus PAM sequences used for genome editing
peaks	peak input file path or a GenomicRanges object that contains genomic regions to be searched for potential offtargets
format	Format of the gRNA and peak input file. Currently, fasta and bed are supported for gRNA and peak input file respectively
peaks.withHeader	Indicate whether the peak input file contains header, default FALSE
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
overlap.gRNA.positions	The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18 for SpCas9.
upstream	upstream offset from the peak start to search for off targets, default 20
downstream	downstream offset from the peak end to search for off targets, default 20
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 protospacer-adjacent motif (PAM), default to any NNN$. Set it to (NAG\|NGG\|NGA)$ if only outputs offtargets with NAG, NGA or NGG PAM
max.mismatch	Maximum mismatch allowed in off target search, default 6
outputDir	the directory where the off target analysis and reports will be written to
allowed.mismatch.PAM	Number of degenerative bases in the PAM.pattern sequence, default to 2
overwrite	overwrite the existing files in the output directory or not, default FALSE
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) for SPcas9 system, which is used in Hsu et al., 2013 cited in the reference section. Please make sure that the number of elements in this vector is the same as the gRNA.size, e.g., pad 0s at the beginning of the vector.
orderOfftargetsBy	criteria to order the offtargets by and the top one will be kept if keepTopOfftargetsOnly is set to TRUE. If set to predicted_cleavage_score (descending order), the offtarget with the highest predicted cleavage score for each peak will be kept. If set to n.mismatch (ascending order), the offtarget with the smallest number of mismatch to the target sequence for each peak will be kept.

`descending` No longer used. In the descending or ascending order. Default to order by predicted cleavage score in descending order and number of mismatch in ascending order When altering orderOfftargetsBy order, please also modify descending accordingly

`keepTopOfftargetsOnly` Output all offtargets or the top offtarget per peak using the orderOfftargetsBy criteria, default to the top offtarget

`scoring.method` Indicates which method to use for offtarget cleavage rate estimation, currently two methods are supported, Hsu-Zhang and CFDscore

`subPAM.activity` Applicable only when scoring.method is set to CFDscore A hash to represent the cleavage rate for each alternative sub PAM sequence relative to preferred PAM sequence

`subPAM.position` Applicable only when scoring.method is set to CFDscore The start and end positions of the sub PAM. Default to 22 and 23 for SP with 20bp gRNA and NGG as preferred PAM

`PAM.location` PAM location relative to gRNA. For example, default to 3prime for spCas9 PAM. Please set to 5prime for cpf1 PAM since it's PAM is located on the 5 prime end

`mismatch.activity.file` Applicable only when scoring.method is set to CFDscore A comma separated (csv) file containing the cleavage rates for all possible types of single nucleotide mismatch at each position of the gRNA. By default, using the supplemental Table 19 from Doench et al., Nature Biotechnology 2016

`n.cores.max` Indicating maximum number of cores to use in multi core mode, i.e., parallel processing, default 1 to disable multicore processing for small dataset.

Value

a tab-delimited file offTargetsInPeakRegions.tsv, containing all input peaks with potential gRNA binding sites, mismatch number and positions, alignment to the input gRNA and predicted cleavage score.

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 Lihua Julie Zhu, Benjamin R. Holmes, Neil Aronin and Michael Brodsky. CRISPRseek: a Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. Plos One Sept 23rd 2014 Lihua Julie Zhu (2015). Overview of guide RNA design tools for CRISPR-Cas9 genome editing technology. Frontiers in Biology August 2015, Volume 10, Issue 4, pp 289-296

See Also

GUIDEseq

Examples

```
#### the following example is also part of annotateOffTargets.Rd
if (interactive())
{
    library("BSgenome.Hsapiens.UCSC.hg19")
    library(GUIDEseq)
    peaks <- system.file("extdata", "T2plus100OffTargets.bed",
        package = "CRISPRseek")
    gRNAs <- system.file("extdata", "T2.fa",
        package = "CRISPRseek")
    outputDir = getwd()
    offTargets <- offTargetAnalysisOfPeakRegions(gRNA = gRNAs, peaks = peaks,
        format=c("fasta", "bed"),
        peaks.withHeader = TRUE, BSgenomeName = Hsapiens,
        upstream = 25L, downstream = 25L, PAM.size = 3L, gRNA.size = 20L,
        orderOfftargetsBy = "predicted_cleavage_score",
        PAM = "NGG", PAM.pattern = "(NGG|NAG|NGA)$", max.mismatch = 2L,
        outputDir = outputDir,
        allowed.mismatch.PAM = 3, overwrite = TRUE
   )
}
```

`offTargetAnalysisWithBulge` *offTarget Analysis With Bulges Allowed Finding offtargets around peaks from GUIDE-seq or around any given genomic regions with bulges allowed in gRNA or the DNA sequence of offTargets when aligning gRNA and DNA sequences.*

Description

offTarget Analysis With Bulges Allowed Finding offtargets around peaks from GUIDE-seq or around any given genomic regions with bulges allowed in gRNA or the DNA sequence of offTargets when aligning gRNA and DNA sequences.

Usage

```
offTargetAnalysisWithBulge(
  gRNA,
  gRNA.name,
  peaks,
  BSgenomeName,
  mat,
  peaks.withHeader = FALSE,
  peaks.format = "bed",
  gapOpening = 1L,
  gapExtension = 3L,
  max.DNA.bulge = 2L,
  max.mismatch = 10L,
  allowed.mismatch.PAM = 2L,
  upstream = 20L,
  downstream = 20L,
```

```
  PAM.size = 3L,
  gRNA.size = 20L,
  PAM = "NGG",
  PAM.pattern = "NNN$",
  PAM.location = "3prime",
  mismatch.activity.file = system.file("extdata",
    "NatureBiot2016SuppTable19DoenchRoot.xlsx", package = "GUIDEseq")
)
```

Arguments

gRNA	a character string containing the gRNA sequence without PAM
gRNA.name	name of the gRNA
peaks	peak input file path or a GenomicRanges object that contains genomic regions to be searched for potential offtargets
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
mat	nucleotideSubstitutionMatrix, which can be created using nucleotideSubstitutionMatrix.
peaks.withHeader	Indicate whether the peak input file contains header, default FALSE
peaks.format	format of the peak file, default to bed file format. Currently, only bed format is supported
gapOpening	Gap opening penalty, default to 1L
gapExtension	Gap extension penalty, default to 3L
max.DNA.bulge	Total number of bulges allowed, including bulges in DNA and gRNA, default to 2L
max.mismatch	Maximum mismatch allowed in off target search, default 10L
allowed.mismatch.PAM	Number of degenerative bases in the PAM.pattern sequence, default to 2L
upstream	upstream offset from the peak start to search for off targets, default 20
downstream	downstream offset from the peak end to search for off targets, default 20
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 protospacer-adjacent motif (PAM), default to any NNN$. Currently, only support NNN$
PAM.location	PAM location relative to gRNA. For example, default to 3prime for spCas9 PAM. Please set to 5prime for cpf1 PAM since it's PAM is located on the 5 prime end
mismatch.activity.file	Applicable only when scoring.method is set to CFDscore A comma separated (csv) file containing the cleavage rates for all possible types of single nucleotide mismatch at each position of the gRNA. By default, using the supplemental Table 19 from Doench et al., Nature Biotechnology 2016

Author(s)

Lihua Julie Zhu

Examples

```
if (interactive()) {
  library(GUIDEseq)
  peaks <- system.file("extdata","1450-chr14-chr2-bulge-test.bed", package = "GUIDEseq")
 mismatch.activity.file <-system.file("extdata", "NatureBiot2016SuppTable19DoenchRoot.xlsx",
    package = "GUIDEseq")

  gRNA <- "TGCTTGGTCGGCACTGATAG"
  gRNA.name <- "Test1450"
  library(BSgenome.Hsapiens.UCSC.hg38)

  temp <- offTargetAnalysisWithBulge(gRNA = gRNA, gRNA.name = gRNA.name,
     peaks = peaks, BSgenomeName = Hsapiens,
     mismatch.activity.file = mismatch.activity.file)
}
```

---

peaks.gr *example cleavage sites*

Description

An example data set containing cleavage sites (peaks) from getPeaks

Format

GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

Value

peaks.gr	GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

Source

http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1695644

Examples

```
data(peaks.gr)
names(peaks.gr)
peaks.gr
```

PEtagAnalysis *Analysis pipeline for PEtag-seq dataset*

Description

A wrapper function that uses the UMI sequence plus the first few bases of each sequence from R1 reads to estimate the starting sequence library, piles up reads with a user defined window and step size, identify the insertion sites (proxy of cleavage sites), merge insertion sites from plus strand and minus strand, followed by off target analysis of extended regions around the identified insertion sites. Detailed information on additional parameters can be found in GUIDEseqAnalysis manual with help(GUIDEseqAnalysis).

Usage

```
PEtagAnalysis(
  alignment.inputfile,
  umi.inputfile,
  BSgenomeName,
  gRNA.file,
  outputDir,
  keepPeaksInBothStrandsOnly = FALSE,
  txdb,
  orgAnn,
  PAM.size = 3L,
  gRNA.size = 20L,
  overlap.gRNA.positions = c(17, 18),
  PAM.location = "3prime",
  PBS.len = 10L,
  HA.len = 7L,
  ...
)
```

Arguments

alignment.inputfile	The alignment file. Currently supports bam and bed output file with CIGAR information. Suggest run the workflow binReads.sh, which sequentially runs barcode binning, adaptor removal, alignment to genome, alignment quality filtering, and bed file conversion. Please download the workflow function and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/binReads/
umi.inputfile	A text file containing at least two columns, one is the read identifier and the other is the UMI or UMI plus the first few bases of R1 reads. Suggest use getUMI.sh to generate this file. Please download the script and its helper scripts at http://mccb.umassmed.edu/GUIDE-seq/getUMI/
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
gRNA.file	gRNA input file path or a DNAStringSet object that contains the target sequence (gRNA plus PAM)

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

`keepPeaksInBothStrandsOnly`
Indicate whether only keep peaks present in both strands as specified by plus.strand.start.gt.minus.stra max.overlap.plusSig.minusSig and distance.threshold. Please see GUIDEseq-Analysis for details of additional parameters. Default to FALSE for any in vitro system, which needs to be set to TRUE for any in vivo system.

`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#___An 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.en for Drosophila and TxDb.Celegans.UCSC.ce6.ensGene for C.elegans

`orgAnn` organism annotation mapping such as org.Hs.egSYMBOL in org.Hs.eg.db package for human

`PAM.size` PAM length, default 3

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

`overlap.gRNA.positions`
The required overlap positions of gRNA and restriction enzyme cut site, default 17 and 18 for SpCas9.

`PAM.location` PAM location relative to gRNA. For example, default to 3prime for spCas9 PAM. Please set to 5prime for cpf1 PAM since it's PAM is located on the 5 prime end

`PBS.len` Primer binding sequence length, default to 10.

`HA.len` Homology arm sequence length, default to 7.

`...` Any parameters in GUIDEseqAnalysis can be used for this function. Please type help(GUIDEseqAnalysis for detailed information.

Value

`offTargets` a data frame, containing all input peaks with potential gRNA binding sites, mismatch number and positions, alignment to the input gRNA, predicted cleavage score, PBS (primer binding sequence), and HAseq (homology arm sequence).

`merged.peaks` merged peaks as GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

`peaks` GRanges with count (peak height), bg (local background), SNratio (signal noise ratio), p-value, and option adjusted p-value

`uniqueCleavages`
Cleavage sites with one site per UMI as GRanges with metadata column total set to 1 for each range

`read.summary` One table per input mapping file that contains the number of reads for each chromosome location

Author(s)

Lihua Julie Zhu

References

Lihua Julie Zhu, Michael Lawrence, Ankit Gupta, Herve Pages, Alper Ku- cukural, Manuel Garber and Scot A. Wolfe. GUIDEseq: a bioconductor package to analyze GUIDE-Seq datasets for CRISPR-Cas nucleases. BMC Genomics. 2017. 18:379

See Also

GUIDEseqAnalysis

Examples

```
if(!interactive())
    {
        library("BSgenome.Hsapiens.UCSC.hg19")
        library(TxDb.Hsapiens.UCSC.hg19.knownGene)
        library(org.Hs.eg.db)
        umiFile <- system.file("extdata", "UMI-HEK293_site4_chr13.txt",
           package = "GUIDEseq")
        alignFile <- system.file("extdata","bowtie2.HEK293_site4_chr13.sort.bam" ,
            package = "GUIDEseq")
        gRNA.file <- system.file("extdata","gRNA.fa", package = "GUIDEseq")
        PET.res <- PEtagAnalysis(
            alignment.inputfile = alignFile,
            umi.inputfile = umiFile,
            gRNA.file = gRNA.file,
            orderOfftargetsBy = "peak_score",
            descending = TRUE,
            keepTopOfftargetsBy = "predicted_cleavage_score",
            scoring.method = "CFDscore",
            BSgenomeName = Hsapiens,
            txdb = TxDb.Hsapiens.UCSC.hg19.knownGene,
            orgAnn = org.Hs.egSYMBOL,
            outputDir = "PEtagTestResults",
            min.reads = 80, n.cores.max = 1,
            keepPeaksInBothStrandsOnly = FALSE,
            PBS.len = 10L,
            HA.len = 7L
            )
        PET.res$offTargets
        names(PET.res)
    }
```

---

plotAlignedOfftargets *Plot offtargets aligned to the target sequence*

Description

Plot offtargets aligned to the target sequence

Usage

```
plotAlignedOfftargets(
  offTargetFile,
  sep = "\t",
  header = TRUE,
  gRNA.size = 20L,
  input.DNA.bulge.symbol = "^",
  input.RNA.bulge.symbol = "-",
```

```
  input.match.symbol = ".",
  plot.DNA.bulge.symbol = "DNA.bulge",
  plot.RNA.bulge.symbol = "-",
  plot.match.symbol = ".",
  color.DNA.bulge = "red",
  size.symbol = 3,
 color.values = c(A = "#B5D33D", T = "#AE9CD6", C = "#6CA2EA", G = "#FED23F", `-` =
    "gray", . = "white"),
  PAM = "GGG",
  body.tile.height = 2.5,
  header.tile.height = 3.6,
  hline.offset = 3.8,
  plot.top.n,
  insertion.score.column = c("n.distinct.UMIs", "peak_score"),
  insertion.score.column.prefix,
  width.IR = 2.5,
  width.RIR = 2.5,
  family = "sans",
  hjust = "middle",
  vjust = 0.5
)
```

Arguments

offTargetFile	The path of the file offTargetsInPeakRegions.xls that stores the offtargets to be plotted. This file is the output file from the function GUIDEseqAnalysis.
sep	Field delimiter for the file specified as offTargetFile, default to tab dilimiter
header	Indicates whether there is header in the file specified as offTargetFile, default to TRUE
gRNA.size	Size of the gRNA, default to 20 for SpCas9 system
input.DNA.bulge.symbol	The symbol used to represent DNA bulges in the file specified as offTargetFile, default to "^"
input.RNA.bulge.symbol	The symbol used to represent RNA bulges in the file specified as offTargetFile, default to "-"
input.match.symbol	The symbol used to represent matched bases in the file specified as offTargetFile, default to "."
plot.DNA.bulge.symbol	The symbol used to represent DNA bulges in the figure to be generated, default to DNA.bulge, i.e., the nucleotide in the DNA bulge. Alternatively, you can specify a symbol to represent all DNA bulges such as "I".
plot.RNA.bulge.symbol	The symbol used to represent RNA bulges in the figure to be generated, default to "-"
plot.match.symbol	The symbol used to represent matched bases in the figure to be generated, default to "."
color.DNA.bulge	The color used to represent DNA bulges in the figure to be generated, default to "red"

size.symbol	The size used to plot the bases, and the symbols of DNA/RNA bulges, default to 3
color.values	The color used to represent different bases, DNA bulges, and RNA bulges.
PAM	PAM sequence in the target site, please update it to the exact PAM sequence in the input target site.
body.tile.height	Specifies the height of each plotting tile around each base/symbol for offtargets, default to 2.5
header.tile.height	Specifies the height of each plotting tile around each base/symbol for the target sequence on the very top, default to 3.6
hline.offset	Specifies the offset from the top border to draw the horizontal line below the gRNA sequence, default to 3.8. Increase it to move the line down and decrease it to move the line up.
plot.top.n	Optional. If not specified, all the offtargets in the input file specified as offTargetFile will be included in the plot. With a very large number of offtargets, users can select the top n offtargets to be included in the plot. For example, set plot.top.n = 20 to include only top 20 offtargets in the plot. Please note offtargets are ordered by the n.distinct.UMIs or peak_score from top to bottom.
insertion.score.column	"n.distinct.UMIs" or "peak_score" to be included on
insertion.score.column.prefix	to designate sample name e.g., S1 which means that two of columns are named as S1.peak_score and S1.n.distinct.UMIs in the input file. Useful if the input file is generated by the function combineOfftargets the right side of the alignment as Insertion Events. Relative Insertion Rate (RIR) divided by ontarget peak_score/n.distinct.UMIs. For example, RIR for ontarget should be 100
width.IR	For adjusting the width of the IR output
width.RIR	For adjusting the width of the RIR output
family	font family, default to sans (Arial). Other options are serif (Times New Roman) and mono (Courier). It is possible to use custom fonts with the extrafont package with the following commands install.packages("extrafont") library(extrafont) font_import() loadfonts(device = "postscript")
hjust	horizontal alignment
vjust	vertical alignment

Value

a ggplot object

Author(s)

Lihua Julie Zhu

Examples

```
offTargetFilePath <- system.file("extdata/forVisualization",
 "offTargetsInPeakRegions.xls",
 package = "GUIDEseq")
fig1 <- plotAlignedOfftargets(offTargetFile = offTargetFilePath,
```

```
    plot.top.n = 20,
    plot.match.symbol = ".",
    plot.RNA.bulge.symbol = "-",
    insertion.score.column = "peak_score")
fig1

fig2 <- plotAlignedOfftargets(offTargetFile = offTargetFilePath,
    plot.top.n = 20,
    plot.match.symbol = ".",
    plot.RNA.bulge.symbol = "-",
    insertion.score.column = "n.distinct.UMIs")
fig2
```

`plotHeatmapOfftargets` *Plot offtargets from multiple samples as heatmap*

Description

Plot offtargets from multiple samples as heatmap

Usage

```
plotHeatmapOfftargets(
  mergedOfftargets,
  min.detection.rate = 0.1,
  font.size = 12,
  on.target.predicted.score = 1,
  IR.normalization = c("sequence.depth", "on.target.score", "sum.score", "none"),
  top.bottom.height.ratio = 3,
  dot.distance.breaks = c(5, 10, 20, 40, 60),
  dot.distance.scaling.factor = c(0.4, 0.6, 0.8, 1.2, 2),
  bottom.start.offset = 8,
  color.low = "white",
  color.high = "blue",
  sample.names,
  insertion.score.column = c("n.distinct.UMIs", "peak_score")
)
```

Arguments

`mergedOfftargets`
: a data frame from running the combineOfftargets function

`min.detection.rate`
: minimum relative detection rate to be included in the heatmap

`font.size`
: font size for x labels and numbers along the y-axis.

`on.target.predicted.score`
: Default to 1 for the CFDscore scoring method. Set it to 100 for the Hsu-Zhang scoring method.

`IR.normalization`
: Default to sequence.depth which uses the sequencing depth for each sample in the input file to calculate the relative insertion rate (RIR). Other options are "on.target.score" and "sum.score" which use the on-target score for each sample and the sum of all on-target and off-target scores to calculate the RIR respectively. The score can be either peak.score or n.distinct.UMIs as specified by the parameter insertion.score.column

`top.bottom.height.ratio`
: the ratio of the height of top panel vs that of the bottom panel.

`dot.distance.breaks`
: a numeric vector for specifying the minimum number of rows in each panel to use the the corresponding distance in dot.distance.scaling.factor between consecutive dots along the y-axis. In the default setting, dot.distance.breaks and dot.distance.scaling.factor are set to c(5, 10, 20, 40, 60) and c(0.4, 0.6, 0.8, 1.2, 2) respectively, which means that if the number of rows in each panel is greater than or equal to 60, 40-59, 20-39, 10-19, 5-9, and less than 5,then the distance between consecutive dots will be plotted 2, 1.2, 0.8, 0,6, 0.4, and 0.2 (half of 0.4) units away in y-axis respectively.

`dot.distance.scaling.factor`
: a numeric vector for specifing the distance between two consecutive dots. See dot.distance.breaks for more information.

`bottom.start.offset`
: Default to 2, means that place the top number in the bottom panel 2 units below the top border. Increase the value will move the number away from the top border.

`color.low` The color used to represent the lowest indel rate, default to white

`color.high` The color used to represent the highest indel rate the intermediate indel rates will be colored using the color between color.low and color.high. Default to blue.

`sample.names` Optional sample Names used to label the x-axis. If not provided, x-axis will be labeled using the sample names provided in the GUIDEseqAnalysis step.

`insertion.score.column`
: "n.distinct.UMI" or "peak_score" to be included on the right side of the alignment as Insertion Events. Relative Insertion Rate (RIR) divided by ontarget peak_score/n.distinct.UMI. For example, RIR for ontarget should be 100

Value

a ggplot object

Author(s)

Lihua Julie Zhu

Examples

```
if (interactive())
{
  mergedOfftargets <-
        read.table(system.file("extdata/forVisualization",
      "mergedOfftargets.txt",
       package = "GUIDEseq"),
                   sep ="\t", header = TRUE)
```

```
 figs <- plotHeatmapOfftargets(mergedOfftargets,
                  min.detection.rate = 2.5,
                  IR.normalization = "on.target.score",
                  top.bottom.height.ratio = 12,
                  bottom.start.offset = 6,
                  dot.distance.scaling.factor = c(0.2,0.2,0.4,0.4, 0.4),
                  sample.names = c("Group1", "Group2"))
                  figs[[1]]/figs[[2]] +
 plot_layout(heights = unit(c(2,1),
                            c('null', 'null')))

figs = plotHeatmapOfftargets(mergedOfftargets,
                 min.detection.rate = 1.2,
                 IR.normalization = "sum.score",
                 top.bottom.height.ratio = 12,
                 bottom.start.offset = 6,
                 dot.distance.scaling.factor = c(0.2,0.2,0.4,0.4, 0.4),
                 sample.names = c("Group1", "Group2"))
                 figs[[1]]/figs[[2]] +
                 plot_layout(heights = unit(c(2,1),
                  c('null', 'null')))
 figs <- plotHeatmapOfftargets(mergedOfftargets,
    min.detection.rate = 0.2,
    IR.normalization = "sequence.depth",
    top.bottom.height.ratio = 12,
    bottom.start.offset = 6,
    dot.distance.scaling.factor = c(0.2,0.2,0.2,0.2, 0.2),
    sample.names = c("Group1", "Group2"))
figs[[1]]/figs[[2]] +
    plot_layout(heights = unit(c(2,1),
    c('null', 'null')))
figs = plotHeatmapOfftargets(mergedOfftargets,
    min.detection.rate = 3,
    IR.normalization = "none",
    top.bottom.height.ratio = 12,
    bottom.start.offset = 6,
    dot.distance.scaling.factor = c(0.2,0.2,0.7,0.7, 0.7),
    sample.names = c("Group1", "Group2"))
    figs[[1]]/figs[[2]]
plot_layout(heights = unit(c(2,1),
                c('null', 'null')))
}
```

`plotTracks` *Plot offtargets as manhantann plots or along all chromosomes with one track per chromosome, or scatter plot for two selected measurements*

Description

Plot offtargets as manhantann plots or along all chromosomes with one track per chromosome, or scatter plot for two selected measurements

Usage

```
plotTracks(
  offTargetFile,
  sep = "\t",
  header = TRUE,
  gRNA.size = 20L,
  PAM.size = 3L,
  cleavage.position = 19L,
  chromosome.order = paste0("chr", c(1:22, "X", "Y", "M")),
  xlab = "Chromosome Size (bp)",
  ylab = "Peak Score",
  score.col = c("peak_score", "n.distinct.UMIs", "total.match", "gRNA.match",
    "total.mismatch.bulge", "gRNA.mismatch.bulge", "predicted_cleavage_score"),
  transformation = c("log10", "none"),
  title = "",
  axis.title.size = 12,
  axis.label.size = 8,
  strip.text.y.size = 9,
  off.target.line.size = 0.6,
  on.target.line.size = 1,
  on.target.score = 1,
  on.target.color = "red",
  off.target.color = "black",
  strip.text.y.angle = 0,
  scale.grid = c("free_x", "fixed", "free", "free_y"),
  plot.type = c("manhattan", "tracks", "scatter"),
  family = "serif",
  x.sep = 6e+06,
  plot.zero.logscale = 1e-08,
  scale.chrom = TRUE
)
```

Arguments

offTargetFile	The file path containing off-targets generated from GUIDEseqAnalysis
sep	The separator in the file, default to tab-delimited
header	Indicates whether the input file contains a header, default to TRUE
gRNA.size	The size of the gRNA, default 20
PAM.size	PAM length, default 3
cleavage.position	the cleavage position of Cas nuclease, default to 19 for SpCas9.
chromosome.order	The chromosome order to plot from top to bottom
xlab	The x-asix label, default to Chromosome Size (bp)
ylab	The y-asix label, default to Peak Score. Change it to be consistent with the score.col
score.col	The column used as y values in the plot. Available choices are peak_score, n.distinct.UMIs, total.match, gRNA.match, total.mismatch.bulge, gRNA.mismatch.bulge, and predicted_cleavage_score. When plot.type is set to scatter, a vector of size

two can be set. Otherwise, a scatter plot with log10 transformed n.distinct.UMIs and log10 transformed predicted_cleavage_score will be plotted.

`transformation` Indicates whether plot the y-value in log10 scale or in the original scale. When scale.col is set to total.match, gRNA.match, total.mismatch.bulge, and gRNA.mismatch.bulge, transformation will not be applied and the data will be plotted in the original scale. When plot.type is set to "scatter", a vector of size two is required when score.col is a vector of size two. Examples are c("log10", "log10"), c("none", "none"), c(log10", "none"), and c("none", "log10").

`title` The figure title, default to none.

`axis.title.size`
The font size for the axis labels, default to 12

`axis.label.size`
The font size for the tick labels, default to 8

`strip.text.y.size`
The font size for the strip labels, default to 9

`off.target.line.size`
The line size to depict the off-targets, default to 0.6

`on.target.line.size`
The line size to depict the on-targets, default to 1

`on.target.score`
The score for the on-target, default to 1 for CFD scoring system. This is the maximum score in the chosen scoring system. Change it accordingly if different off-target scoring system is used.

`on.target.color`
The line color to depict the on-targets, default to red

`off.target.color`
The line color to depict the off-targets, default to black

`strip.text.y.angle`
The angel for the y strip text, default to 0. Set it to 45 if angled representation is desired

`scale.grid` Used to set the scales in facet_grid, default to free_x, meaning that scales vary across different x-axis, but fixed in y-axis. Other options are fixed, free, and free_y meaning that scales shared across all facets, vary across both x- and y-axises, and vary across y-axis only, respectively. For details, please type ?ggplot2::facet_grid

`plot.type` Plot type as tracks by individual chromosome or manhattan plot with all chromosome in one plot

`family` font family, default to sans (Arial). Other options are serif (Times New Roman) and mono (Courier). It is possible to use custom fonts with the extrafont package with the following commands install.packages("extrafont") library(extrafont) font_import() loadfonts(device = "postscript")

`x.sep` For transforming the x-axis to allow sufficient spaces between small chromosomes default to 6000000

`plot.zero.logscale`
Specifying "none" to filter out score.col with zeros when plotting in log10 scale. Specify a very small numeric number if you intend to show the zeros in log scale in the figure. If users specify a number that's bigger than any positive score, plot.zero.logscale will be set to the minimum positive score divided by 10.

`scale.chrom` Applicable to manhatann plot only. TRUE or FALSE default to TRUE to space offtargets evenly along x-axis.

Value

a ggplot object

Author(s)

Lihua Julie Zhu

Examples

```
if (interactive())
{
   offTargetFilePath <- system.file("extdata/forVisualization",
      "offTargetsInPeakRegions.xls",
       package = "GUIDEseq")
  fig1 <- plotTracks(offTargetFile = offTargetFilePath)
  fig1
  fig2 <- plotTracks(offTargetFile = offTargetFilePath,
    score.col = "total.mismatch.bulge",
    ylab = "Total Number of Mismatches and Bulges")
  fig2
  fig3 <- plotTracks(offTargetFile = offTargetFilePath,
     score.col = "total.match",
     ylab = "Total Number of Matches")
  fig3
  fig4 <- plotTracks(offTargetFile = offTargetFilePath,
      score.col = "gRNA.match",
      ylab = "Number of Matches in gRNA")
  fig4
  fig5 <- plotTracks(offTargetFile = offTargetFilePath,
      score.col = "gRNA.mismatch.bulge",
      ylab = "Number of Mismatches and Bulges in gRNA")
  fig5
  fig6 <- plotTracks(offTargetFile = offTargetFilePath,
     score.col = "predicted_cleavage_score",
     ylab = "CFD Score",
     scale.grid = "fixed",
     transformation = "none")
 fig6

 ## manhattan plot
  fig <- plotTracks(offTargetFile = offTargetFilePath,
        score.col = "total.mismatch.bulge", axis.title.size =9,
        plot.type =  "manhattan",
        ylab = "Number of Mismatches and Bulges in gRNA Plus PAM")
   fig
  fig <- plotTracks(offTargetFile = offTargetFilePath,
       score.col = "total.match", axis.title.size =9,
       plot.type =  "manhattan",
       ylab = "Number of Matches in gRNA Plus PAM")
   fig
fig <- plotTracks(offTargetFile = offTargetFilePath,
                 score.col = "gRNA.match",axis.title.size =9,
                 plot.type =  "manhattan",
                 ylab = "Number of Matches in gRNA")
fig
fig <- plotTracks(offTargetFile = offTargetFilePath,
```

```
                score.col = "gRNA.mismatch.bulge", axis.title.size =9,
                plot.type =  "manhattan",
                ylab = "Number of Mismatches and Bulges in gRNA")
fig

plotTracks(offTargetFile = offTargetFilePath,
    #'score.col = "predicted_cleavage_score",
    axis.title.size =9, family = "serif", plot.zero.logscale = 1e-6,
    plot.type =  "manhattan", transformation = "log10",
    ylab = "CFD Score")

plotTracks(offTargetFile = offTargetFilePath,
     score.col = "peak_score",
     axis.title.size =9,
     plot.type =  "manhattan",
     ylab = "Number of Insertion Events")

plotTracks(offTargetFile = offTargetFilePath,
     score.col = "n.distinct.UMIs",
     axis.title.size =9,
     plot.type =  "manhattan",
     ylab = "Number of Insertion Events")

# default scatter plot with blue line from fitting the entire dataset
# and the red line from fitting the subset with CFD score > 0
plotTracks(offTargetFile = offTargetFilePath,
    axis.title.size =9, plot.zero.logscale = 1e-8,
    plot.type =  "scatter")

# select the x, y, the transformation of x and y,
# and the labels on the scatter plot

plotTracks(offTargetFile = offTargetFilePath,
    axis.title.size =9,
    score.col = c("n.distinct.UMIs", "predicted_cleavage_score"),
    transformation = c("log10", "log10"),
    plot.type =  "scatter", plot.zero.logscale = 1e-8,
    xlab = "log10(Number of Insertion Events)",
    ylab = "log10(CFD score)")

}
```

`uniqueCleavageEvents` *example unique cleavage sites*

Description

An example data set containing cleavage sites with unique UMI, generated from getUniqueCleavageEvents

Value

cleavage.gr Cleavage sites with one site per UMI as GRanges with metadata column total set to 1 for each range

unique.umi.plus.R2 a data frame containing unique cleavage site from R2 reads mapped to plus strand with the following columns chr.y (chromosome of readSide.y/R2 read) chr.x (chromosome of readSide.x/R1 read) strand.y (strand of readSide.y/R2 read) strand.x (strand of readSide.x/R1 read) start.y (start of readSide.y/R2 read) end.x (start of readSide.x/R1 read) UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

unique.umi.minus.R2 a data frame containing unique cleavage site from R2 reads mapped to minus strand with the following columns chr.y (chromosome of readSide.y/R2 read) chr.x (chromosome of readSide.x/R1 read) strand.y (strand of readSide.y/R2 read) strand.x (strand of readSide.x/R1 read) end.y (end of readSide.y/R2 read) start.x (start of readSide.x/R1 read) UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

unique.umi.plus.R1 a data frame containing unique cleavage site from R1 reads mapped to minus strand without corresponding R2 reads mapped to the plus strand, with the following columns chr.y (chromosome of readSide.y/R2 read) chr.x (chromosome of readSide.x/R1 read) strand.y (strand of readSide.y/R2 read) strand.x (strand of readSide.x/R1 read) start.x (start of readSide.x/R1 read) start.y (start of readSide.y/R2 read) UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

unique.umi.minus.R1 a data frame containing unique cleavage site from R1 reads mapped to plus strand without corresponding R2 reads mapped to the minus strand, with the following columns chr.y (chromosome of readSide.y/R2 read) chr.x (chromosome of readSide.x/R1 read) strand.y (strand of readSide.y/R2 read) strand.x (strand of readSide.x/R1 read) end.x (end of readSide.x/R1 read) end.y (end of readSide.y/R2 read) UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

all.umi a data frame containing all the mapped reads with the following columns. readName (read ID), chr.x (chromosome of readSide.x/R1 read), start.x (start of eadSide.x/R1 read), end.x (end of eadSide.x/R1 read), mapping.qual.x (mapping quality of readSide.x/R1 read), strand.x (strand of readSide.x/R1 read), cigar.x (CIGAR of readSide.x/R1 read) , readSide.x (1/R1), chr.y (chromosome of readSide.y/R2 read) start.y (start of readSide.y/R2 read), end.y (end of readSide.y/R2 read), mapping.qual.y (mapping quality of readSide.y/R2 read), strand.y (strand of readSide.y/R2 read), cigar.y (CIGAR of readSide.y/R2 read), readSide.y (2/R2) R1.base.kept (retained R1 length), R2.base.kept (retained R2 length), distance (distance between mapped R1 and R2), UMI (unique molecular identifier (umi) or umi with the first few bases of R1 read)

Source

http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1695644

Examples

```
data(uniqueCleavageEvents)
names(uniqueCleavageEvents)
sapply(uniqueCleavageEvents, class)
uniqueCleavageEvents[[1]]  # GRanges object
lapply(uniqueCleavageEvents, dim)
```

Index